Algorithmic and architectural constraints on human 3D visual inferences

Tyler Bonnen
UC Berkeley

Anthony D. Wagner
Stanford University

Daniel L.K. Yamins
Stanford University

ABSTRACT

Perception unfolds across multiple timescales. Ventral temporal cortex (VTC) supports visual inferences that are possible 'at a glance' (i.e.<200ms), such as object classification. Other visual inferences, such as inferring the 3D shape of unfamiliar objects, require more time. Using a combination of psychophysics, electrophysiological, and lesion data, here we identify neural structures and algorithms that underlie this ability. First, we compare an online cohort of human participants to electrophysiological recordings from macaque VTC. While performance 'at a glance' is predicted by VTC responses, humans outperform VTC with increased stimulus viewing time. Next, we demonstrate that a neural system downstream of VTC, medial temporal cortex (MTC), plays a causal role in these temporally extended visual inferences. Finally, through a series of in lab eyetracking experiments, we demonstrate that sequential visual sampling of object features is both reliable across participants and necessary for performance. From these data, we suggest that MTC support visual inferences by integrating over visuospatial sequences, providing algorithmic and architectural constraints for theories and models of human perception.

Keywords: ventral temporal cortex; medial temporal cortex; object perception; 3D vision; temporal dynamics

INTRODUCTION

There is temporal structure in how we perceive the world. Many visual attributes can be inferred 'at a glance' (Potter, 1975), an ability which depends on ventral temporal cortex (VTC) (DiCarlo, Zoccolan, & Rust, 2012). However, not all visual inferences are possible 'at a glance' (Findlay & Gilchrist, 2003). This is due, in part, to constraints on the primate visual system (Van Essen & Anderson, 1990): high-acuity visual information is only maintained at the central visual field, i.e., the fovea (Hirsch & Curcio, 1989), and so to collect precise visual information from the environment we shift the location of our gaze roughly three times per second (Liversedge & Findlay, 2000). Such visual 'routines' are thought to underlie many visual abilities (Ullman, 1987), including inferring the shape of unfamiliar objects. Medial temporal cortex (MTC), downstream of VTC, has been proposed as a neuroanatomical substrate that supports visual inferences not possible from VTC alone (Murray & Bussey, 1999; Bussey & Saksida, 2002). Lesions to MTC result in profound impairments in tasks designed to assay 'complex' visual object perception, such as inferring the 3D shape of objects (Barense, Gaffan, & Graham, 2007). Recent computational work demonstrates that while MTC-lesioned human performance resembles computation proxies for VTC (e.g., convolutional neural networks, CNNs), MTC-intact participants radically exceed MTC-lesioned/CNN performance (Bonnen, Yamins, & Wagner, 2021). It is possible that MTC supports these visual inferences by integrating over sequences of visual information. Here we evaluate this claim through a series of experiments that integrate psychophysics, electrophysiological, and lesion data. We use variations of two experimental designs ('oddity' and 'match to sample'). We draw from two datasets: one that has stimuli associated with electrophysiological recordings from VTC in macaques (Majaj, Hong, Solomon, & DiCarlo, 2015), and another that has stimuli associated with MTC-lesioned human participants (Barense et al., 2007). We administer these stimuli to human participants through a series of online and in-lab experiments. We also compare human performance to the accuracy supported by computational proxies for VTC (e.g., CNNs). These complementary experimental designs and stimulus sets enable us to investigate the causal roles that time, eye movements, and MTC play in human shape inferences.

RESULTS

Humans outperform VTC and CNNs with sufficient time. We first compare human performance directly to the accuracy supported by a linear readout of ventral temporal cortex (i.e., VTC-supported performance) and convolutional neural networks (Imagenet-optimized CNNs). Using stimuli and ele-

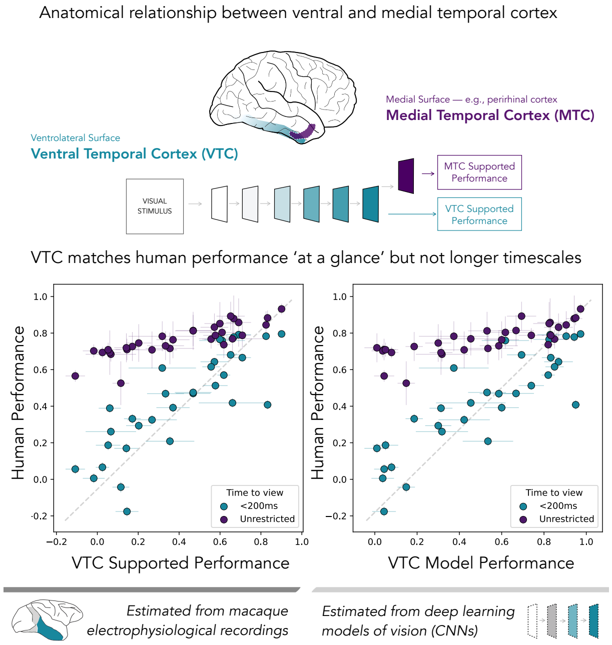

Figure 1: VTC (left) and CNNs (right) match human performance 'at a glance' (green) but not longer timescales (purple).

ctrophysiological recordings previously collected from macaque inferior temporal cortex (Majaj et al., 2015), we evaluate human performance via two online experimental protocols. We compare time restricted human accuracy to VTC-supported performance using a zero-delay two alternative forced choice task. Each trial is initiated by the participant; a stimulus is presented and disappears after 100ms, followed by a full-screen white noise mask, and then a two alternative forced choice cuing the participant to "choose the image that contains the object shown in the previous screen." We observe a striking correspondence between time restricted human performance and a linear readout of the VTC on this zero-delay match to sample paradigm (Fig 1 (green); $\beta = 0.87$, $F(1,30)$ = $8.34, P = 3$x10^{-9}). Human accuracy and VTC-supported performance did not significantly differ (paired ttest $\beta = -0.02$, $t(31)$ = $-0.85, P = 0.40$). We administer these same stimuli via an 'oddity' task to participants ($N = 297$) how are presented with three images and instructed to identify the object which is different from the others. Here, human participants outperform a linear readout of VTC (Fig 1 (purple): paired ttest $\beta = .24$, $t(31) = 9.50$, $P =$ 1x10^{-10}). As a control, we validate that time unrestricted participants (n=50) participants in match-to-sample experiment exceed the performance of time restricted participants (paired ttest $t(30)$ = $12.01, P = 5.44$ x 10^{-13}). Remarkably, we find that a computational proxy for VTC (i.e., a task optimized CNN) demonstrates the same pattern as electrophysiological recordings (Fig 1; right), predicting time restricted humans (ols regression $\beta = .81$, $F(1,30) = 13.33$, $P =$ 4x10^{-14}) while being outperformed by time restricted participants (paired ttest $\beta = .16$, $t(31) = 5.38$, $P =$ 7x10^{-6}).

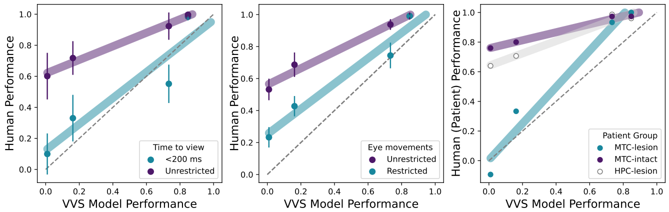

Figure 2: Temporally extended visual inferences (left) rely on eye movements (middle) and medial temporal cortex (right)

MTC supports temporally extended visual inferences. Here we draw from a previously collected dataset (Barense et al., 2007) that compared MTC-lesioned to MTC-intact human participants on visual discrimination tasks. We administer four conditions in this dataset to human participants, enabling us to compare time-restricted/-unrestricted performance directly to MTC-lesioned performance. First, we find that time restricted human performance is predicted by VTC model performance ($\beta = 0.57$, $F(1,195)$ = $16.92, P = 4$ x 10^{-40}; Fig. 2 left, green) while time unrestricted performance exceeds time restricted performance (unpaired ttest between the average performance of time restricted/unrestricted condition-level accuracy: $t(327)$ = $5.67, P = 3.04$ x 10^{-8}; Fig. 2 left, purple). As a control, we administer a 'self-paced' match-to-sample task where participants are given unrestricted time to encode the stimulus: time unrestricted participants (n=20) outperform time restricted participants (paired ttest $t(275)$ = $4.56, P = 7.17$ x 10^{-6}). That is, the temporal pattern of human visual inferences, previously observed on stimuli from Majaj et al. (2015), is also evident in these experiments using stimuli from Barense et al. (2007). Remarkably, MTC-lesioned performance resembles time restricted human performance, while MTC-intact performance exceeds it (Fig. 2 right), suggesting that MTC plays a causal role in these temporally extended visual inferences.

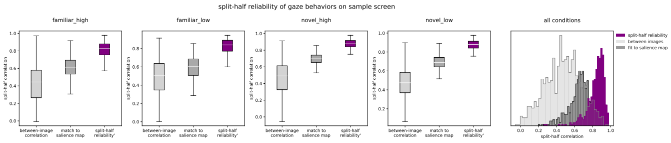

Figure 3: Gaze dynamics are reliable across participants.

Gaze dynamics are necessary/reliable for performance. Here we determine whether MTC-dependent visual inferences require longer viewing times, per se, or depend on a visual 'routine' (i.e., visuospatial integration) using a series of in-lab eye tracking experiments ($N = 97$). First, participants viewed each stimulus on the sample screen at their own pace, but their gaze is restricted to the screen's center (i.e., time unrestricted but gaze restricted). On the match screen, participants are free to move their eyes. Gaze restricted participants are significantly impaired on this task ($t(106)$ = $-4.13, P = 7.14$ x 10^{-5}), suggesting a causal role for these gaze dynamics. Finally, we determine whether these dynamics are reliable across participants using gaze unrestricted participants. We determine the correlation between (random split-half) salience maps associated with each image, across participants. As an empirical null, we use this same score between different images within the same trial, and also compare to bottom-up salience of each image using standard image processing pipelines (Itti, Koch, & Niebur, 1998). Gaze behaviors are reliable across participants, are significantly greater that the empirical null (unpaired ttest $t(1258)$ = $35.97, P = 1.90$ x 10^{-195}; Fig. 3 grey, all panels) as well as the fit to bottom-up salience maps (unpaired ttest $t(835)$ = $26.38, P = 4.58$ x 10^{-112}; Fig. 3 black, all panels), indicating that these gaze behaviors are reliable and not simply driven by bottom-up salience.

CONCLUSION

Our work characterized neural structures (MTC) and algorithms (visuospatial integration) that enable humans to represent the underlying shapes of novel objects. When humans lack either of these (because of damage to MTC or time/gaze restricted viewing) our behaviors are predicted by a linear readout of VTC/CNNs. Taken together, these data offer implementation-level details for longstanding theories of visual integration (Ullman, 1987), as well as architectural and algorithmic constraints on future work that aims to model human 3D visual inferences within a biologically plausible framework.

References

Barense, M. D., Gaffan, D., & Graham, K. S. (2007). The human medial temporal lobe processes online representations of complex objects. *Neuropsychologia*, *45*, 2963–2974. Retrieved from 10.1016/j.neuropsychologia.2007.05.023

Bonnen, T., Yamins, D. L., & Wagner, A. D. (2021). When the ventral visual stream is not enough: A deep learning account of medial temporal lobe involvement in perception. *Neuron*, *109*(17), 2755–2766.

Bussey, T. J., & Saksida, L. M. (2002). The organization of visual object representations: a connectionist model of effects of lesions in perirhinal cortex. *Eur. J. Neurosci.*, *15*(2), 355–364. Retrieved from 10.1046/j.0953-816x.2001.01850.x

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434. Retrieved from 10.1016/j.neuron.2012.01.010

Findlay, J. M., & Gilchrist, I. D. (2003). *Active vision: The psychology of looking and seeing* (No. 37). Oxford University Press.

Hirsch, J., & Curcio, C. A. (1989). The spatial resolution capacity of human foveal retina. *Vision research*, *29*(9), 1095–1101.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, *20*(11), 1254–1259.

Liversedge, S. P., & Findlay, J. M. (2000). Saccadic eye movements and cognition. *Trends in cognitive sciences*, *4*(1), 6–14.

Majaj, N. J., Hong, H., Solomon, E. A., & DiCarlo, J. J. (2015). Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *J. Neurosci.*, *35*(39), 13402–13418. Retrieved from 10.1523/JNEUROSCI.5181-14.2015

Murray, E. A., & Bussey, T. J. (1999). Perceptual–mnemonic functions of the perirhinal cortex. *Trends. Cogn. Sci.*, *3*(4), 142–151. Retrieved from 10.1016/s1364-6613(99)01303-0

Potter, M. C. (1975). Meaning in visual search. *Science*, *187*(4180), 965–966.

Ullman, S. (1987). Visual routines. In *Readings in computer vision* (pp. 298–328). Elsevier.

Van Essen, D. C., & Anderson, C. H. (1990). Information processing strategies and pathways in the primate retina and visual cortex. In *An introduction to neural and electronic networks* (pp. 43–72).