Mid-level Feature Differences Support Early Animacy and Object Size Distinctions: Evidence from Electroencephalography Decoding

Ruosi Wang, Daniel Janini, and Talia Konkle

Abstract

■ Responses to visually presented objects along the cortical surface of the human brain have a large-scale organization reflecting the broad categorical divisions of animacy and object size. Emerging evidence indicates that this topographical organization is supported by differences between objects in mid-level perceptual features. With regard to the timing of neural responses, images of objects quickly evoke neural responses with decodable information about animacy and object size, but are mid-level features sufficient to evoke these rapid neural responses? Or is slower iterative neural processing required to untangle information about animacy and object size from mid-level features, requiring hundreds of milliseconds more processing time? To answer this question, we used EEG to measure human neural responses to images of objects and their texform counterparts—unrecognizable images that preserve some mid-level feature information about texture and coarse form. We found that texform images evoked neural responses with early decodable information about both animacy and real-world size, as early as responses evoked by original images. Furthermore, successful cross-decoding indicates that both texform and original images evoke information about animacy and size through a common underlying neural basis. Broadly, these results indicate that the visual system contains a mid-level feature bank carrying linearly decodable information on animacy and size, which can be rapidly activated without requiring explicit recognition or protracted temporal processing. ■

INTRODUCTION

The ventral visual stream contains extensive information about different object categories, with a large-scale spatial organization of response preferences characterized by the broad categories of animacy and object size (Thorat, Proklova, & Peelen, 2019; Julian, Ryan, & Epstein, 2017; Grill-Spector & Weiner, 2014; Konkle & Caramazza, 2013; Konkle & Oliva, 2012). Classic understanding of the ventral stream posits a hierarchical series of processing stages, en route to a more conceptual format that ultimately abstracts away from perceptual information (Proklova, Kaiser, & Peelen, 2016; Mahon, Anzellotti, Schwarzbach, Zampini, & Caramazza, 2009; e.g., for a review, see Peelen & Downing, 2017). However, emerging evidence has revealed that the broad categorical distinctions of the ventral stream are supported by more primitive perceptual differences among "mid-level features" of texture, shape, and curvature (Jagadeesh & Gardner, 2022; Vinken, Konkle, & Livingstone, 2022; Bao, She, McGill, & Tsao, 2020; Yue, Robert, & Ungerleider, 2020; Jozwik, Kriegeskorte, & Mur, 2016; Long, Yu, & Konkle, 2018; Long, Störmer, & Alvarez, 2017; Long, Konkle, Cohen, & Alvarez, 2016; Baldassi et al., 2013). On this emerging account of visual system processing, the ventral stream represents objects in a rich mid-level feature bank, from which more categorical distinctions can be extracted (e.g., with linear read-out).

Evidence for this mid-level feature bank account comes from recent work by Long et al. (2018) investigating brain responses to a new stimulus class called "texforms" (Long et al., 2016, 2017, 2018; Figure 1A). Texform images are created using a texture-synthesis algorithm (Freeman & Simoncelli, 2011), which preserves some mid-level feature information related to the texture and coarse form of the original depicted objects, while obscuring higher-level shape features like clear contours and explicit shape information. Empirically, people cannot identify what these are at the basic level (e.g., as a "cat"). Long et al. (2018) found that texform images evoked extensive responses along the entire ventral visual cortex with a similar large-scale organization as evoked by original, recognizable images. For example, zones of cortex responding more strongly to original animals also responded more to texformed animals. However, given that fMRI data obscure temporal information, there are a number of possible accounts of these large-scale activations. Thus, in the present study, we examined the time-evolving signatures of visual system processing to ask when there is information about animacy and size in neural responses to texform images relative to their original counterparts.

According to the mid-level feature bank account, rapid feedforward activations of the ventral stream reflect sensitivity to mid-level featural distinctions, which directly carry information about animacy and object size. A strong

Harvard University

Journal of Cognitive Neuroscience 34:9, pp. 1670–1680
https://doi.org/10.1162/jocn_a_01883

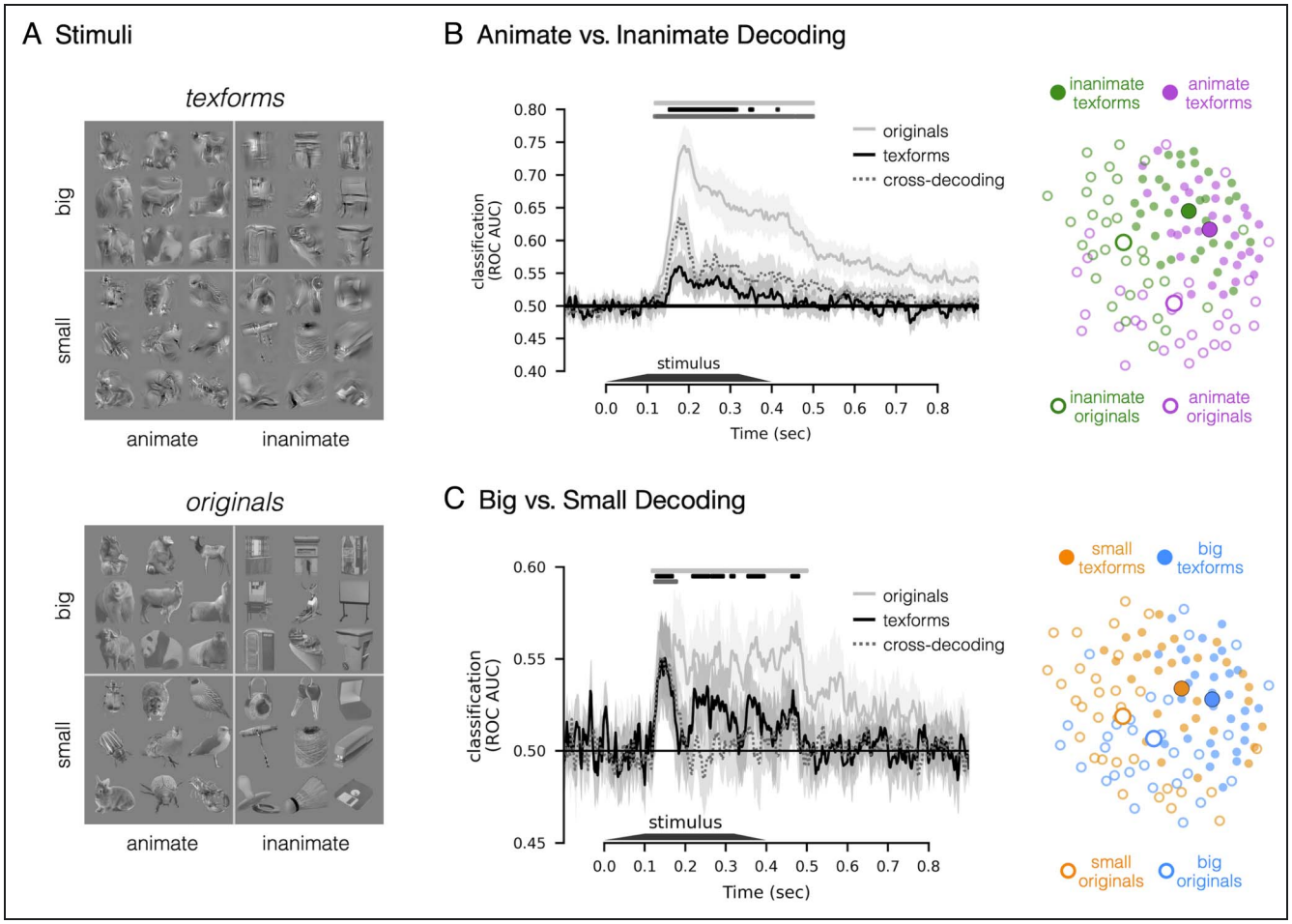

Figure 1. Stimuli and decoding results. (A) Example stimulus images. Each of the four conditions (animacy × size) included 15 exemplars, yielding 60 unrecognizable texforms (upper) and their 60 original counterparts (lower). (B) Time course of animate versus inanimate decoding. Classification accuracy is plotted along the *y* axis, as a function of time (*x* axis), for original (solid sliver lines), texform (solid black lines), and texform-to-original cross-decoding (dashed gray lines). Significant time points are depicted with horizontal lines above the time courses in the corresponding color ($ps < .05$, one-sided signed-rank test, FDR corrected in the time window of interest, 100–500 msec). The shaded region indicates a 95% confidence interval. Adjacent to this axis is a MDS visualization, with a 2-D projection of the pairwise distances in the neural responses to each image from the peak animacy cross-decoding time (176 msec). (C) Time course for big versus small decoding, as in (B). Adjacent MDS plot reflects a 2-D projection of the neural similarity structure measured at the peak size cross-decoding time (140 msec).

temporal prediction of this account is that animacy and object size information emerge early in the time-evolving responses, with comparable timing for texform and original formats. Indeed, EEG and magnetoencephalography decoding studies measuring responses to intact pictures have found that information can be decoded relatively early in the time course of processing about depicted animals versus inanimate objects (Ritchie et al., 2021; Grootswagers, Ritchie, Wardle, Heathcote, & Carlson, 2017; Kaneshiro, Perreau Guimaraes, Kim, Norcia, & Suppes, 2015; Cichy, Pantazis, & Oliva, 2014; Carlson, Tovar, Alink, & Kriegeskorte, 2013) and about big versus small objects (depicted at the same visual size on the screen; Khaligh-Razavi, Cichy, Pantazis, & Oliva, 2018). Furthermore, neurophysiological studies in nonhuman primates also have found that within 100 msec of stimulus onset, information about the animacy of the presented images can be decoded from the population structure of neural responses in V4 and IT (Cauchoix, Crouzet, Fize, & Serre, 2016). Early decoding performance of these high-level properties in original images is consistent with a more primitive underlying format—although this inference is not required by the data.

An alternate temporal prediction is that neural responses to texforms will show more gradual emergence of animacy and object size information, increasing steadily over hundreds of milliseconds. This pattern of data might emerge if texforms contain only very subtle feature differences related to animacy and object size, which are not linearly decodable in an initial feed-forward pass. These subtle differences may trigger later stages of processing, which can reformat and amplify the visual input through more iterative processing steps, so that animacy and object size information is evident in the structure of the responses at later time points. Indeed, Grootswagers, Robinson, Shatek, and Carlson (2019) recently argued for this possibility. They measured responses to texform and original images with EEG, using a rapid presentation

design (Grootswagers, Robinson, & Carlson, 2019) in which they varied the presentation speed of the stimuli. Considering neural responses to original images, they found that animacy and size information could be robustly decoded with presentation rates up to 30 Hz. However, considering neural responses to texform images, they found that animacy could only be decoded at the slowest rate (5 Hz), and size information was not decodable at all. Based on these results, they argued that texforms can elicit animacy signatures, but only given sufficient processing time, and that perhaps higher-order visual areas are required to further "untangle" these features into linearly separable categorical organizations (DiCarlo & Cox, 2007).

Here, we also measured EEG responses to both original and texform images depicting animate and inanimate objects of big and small real-world sizes. However, we used a standard event-related paradigm, allowing us to probe the structure of the neural responses without additional effects of forward and backward masking. To anticipate, we found that both animacy and size information could be decoded from EEG responses to texforms, as early in the time-evolving responses as evoked by original recognizable images. Moreover, we found that classifiers trained on neural responses to texform images were able to predict the animacy and size of responses to original images, indicating that these two image formats reflect animacy and object size information through a common representational basis. Broadly, our results thus support the view that mid-level feature differences contain signatures of animacy and object size, which are available early in the visual processing stream.

METHODS

The experimental data and code used in this study can be found at osf.io/mxrge.

Participants

Participants (n = 19) with normal or corrected-to-normal vision were recruited at the Harvard University community (mean age = 27.5 years, range: 20–42 years; 13 women; one left-handed). This sample size was decided by previous similar studies using EEG decoding (Bae & Luck, 2018; Grootswagers, Ritchie, et al., 2017). All participants provided informed consent and received course credits or financial compensation. We excluded one participant from further analyses because of excessive movements and self-reports of discomfort during the experiment. All procedures were approved by the institutional review board at Harvard University.

Stimuli and Tasks

The stimulus set consisted of 120 total images with 60 recognizable images of 15 big animals, 15 big objects, 15 small animals, and 15 small objects and their texform counterparts (Figure 1A), which were created by undergoing a modified texture-synthesis process (Freeman & Simoncelli, 2011). See Long et al. (2018) for detailed descriptions of stimulus generation. The image set reflects a stratified randomly selected subset of the full stimulus set of Long et al. (2018), which consisted of 240 images.[1]

Stimuli were presented on a 13-in. LCD monitor (1024 × 768 pixels; refresh rate = 60 Hz) at a viewing distance of around 60 cm with a visual angle of 12°, using MATLAB and Psychophysics Toolbox extensions (Brainard, 1997). A bullseye-like fixation remained present at the center of the screen at all times. At the start of each trial, an image was shown for a 400-msec stimulus presentation. In the first 100 msec, the image was linearly faded in, and in the last 83.3 msec, the image was linearly faded out. We made this choice based on the reasoning that it might reduce the abrupt onset and offset signals that might swamp out signatures of later-stage processing. At image offset, there was a 600-msec blank period before the subsequent trial began. We instructed the participants to view the stimulus images attentively while undergoing EEG recording. To minimize artifacts, we included a 1.5-sec "blinking period" every five trials. During this period, the fixation dot turned green to signal the participants that they were encouraged to blink. They were asked to refrain from blinking for the rest of the time.

For each run, all 60 exemplars within a given stimulus type (original or texform) were shown in randomized order and repeated 4 times, resulting in 240 trials (5.32 min). Participants first completed six runs of this protocol in which they saw texform stimuli, followed by six runs with original stimuli. The texform runs were all completed first (rather than alternating with original runs), because we wanted to avoid the possibility that participants hypothesized and looked for correspondences between original images and texform images. This texform-first procedure was also used in the fMRI design from Long et al. (2018).

EEG Recording and Preprocessing

Continuous EEG was recorded from 32 Ag/AgCI electrodes mounted on an elastic cap (EasyCap) and amplified by a Brain Products ActiCHamp system (Brain Vision).[2] The following scalps sites were used: FP1, FP2, F3, F4, FC1, FC2, Cz, C3, C4, CP1, CP2, CP5, CP6, P3, P4, P7, P8, POz, PO3, PO4, PO7, PO8, Oz, O1, O2, Iz, I3, I4. This montage was arranged according to the 10–10 system with some modifications. Specifically, three frontal electrodes were rearranged to have more electrodes over the posterior occipital pole (Stormer, Alvarez, & Cavanagh, 2014). Another two sites, T7 and T8, were also obtained but not used because of the noisy data. The horizontal electrooculogram was measured using electrodes positioned at the external ocular canthi to monitor horizontal eye movements. The vertical electrooculogram was measured at

electrode FP1 to detect eye blinks. All scalp electrodes were on-line referenced to the average of both mastoids and digitized at a rate of 500 Hz.

We conducted EEG data preprocessing and analysis using the MNE-Python package (Gramfort et al., 2014). First, portions of EEG containing excessive muscle movements were identified by visual inspection and removed. Continuous signals were then bandpass filtered with cutoff frequencies of 0.01 Hz and 100 Hz. In the next step, we applied independent component analysis (ICA) for each participant to identify and remove components associated with eye blinks or horizontal eye movements. The ICA-corrected data were segmented into 1000-msec epochs from −100 to 900 msec relative to the stimulus onset and baselined to prestimulus periods. Finally, automated artifact rejection was employed to drop and repair bad epochs using the code package Autoreject (Jas, Engemann, Bekhti, Raimondo, & Gramfort, 2017) with default parameters.

Following these preprocessing steps, participants had, on average, 1373 trials (SD = 77) for texform stimuli and 1358 trials (SD = 85) for original stimuli, with no significant difference between these two stimulus types, $t(17) = .97$, $p = .35$, paired t test. The number of trials did not differ across conditions (big animals, big objects, small animals, and small objects) for either original stimuli, $F(3, 51) = 0.96$, $p = .42$, ANOVA, or texform stimuli, $F(3, 51) = 1.12$, $p = .35$, ANOVA. We also conducted the main analyses without ICA and autoreject procedures in place and obtained the same patterns of results.

Decoding Analyses

Category-level Decoding

A linear discriminant analysis classifier was trained to discriminate animate versus inanimate objects based on neural activation patterns across scalp electrodes, at each time point. The classifier was implemented with *scikit-learn* (Pedregosa et al., 2011) with default parameters (solver: singular value decomposition with threshold of 1.0e-4).

We conducted decoding analyses on supertrials averaged across multiple trials rather than on single-trial data. This procedure is included because previous studies showed that averaging across several trials can improve the signal-to-noise ratio (Bae & Luck, 2018; Grootswagers, Wardle, & Carlson, 2017; Isik, Meyers, Leibo, & Poggio, 2014). In particular, six supertrials were computed for each stimulus exemplar by averaging over two to four trials because the numbers of trials varied across different stimuli after automatic artifact rejection. The number of averaged trials was determined by the recommendation of Grootswagers, Wardle, et al. (2017). This procedure yielded 360 supertrials for recognizable stimuli (e.g., 180 animate / 180 inanimate) and 360 supertrials for texform stimuli. In addition, we also conducted data analysis without applying supertrial averaging and observed the same pattern of results.

Following standard EEG decoding practices on category decoding (Grootswagers, Ritchie, et al., 2017; Carlson et al., 2013; see Grootswagers, Wardle, et al., 2017, for a method review), we employed independent exemplar cross-validation (five-fold), which requires the classifier to generalize to new stimuli. In each fold, the supertrials for 24 animate stimuli and 24 inanimate stimuli (80% exemplars) were used to train the classifier, which was then tested on the supertrials from the remaining six animate stimuli and six inanimate stimuli (20% exemplars). For each fold, we measured the area under the curve of the receiver-operating characteristic (AUC ROC), which reflects an aggregate measure of performance across all possible classification thresholds. Size decoding was computed with a similar logic. Classifiers were trained to discriminate between 24 big and 24 small stimuli and tested on the remaining six big and six small stimuli. In a further analysis to explore tripartite representation (Konkle & Caramazza, 2013), we conducted size decoding separately for big versus small animals and for big versus small inanimate objects.

To ensure the robustness of this AUC ROC estimate, we iterated the above procedure 20 times to minimize the idiosyncrasies in supertrial averaging and five-fold stratified splits. After completing all iterations of cross-validation, the final decoding performance was computed as the average of the 100 decoding attempts (5 folds × 20 iterations).

Cross-decoding

A similar decoding procedure was followed for the cross-decoding analyses but trained on one stimulus type and tested it on the other. For example, in one-fold, the classifier was trained using supertrials from 24 animate exemplars and 24 inanimate exemplars in their texform format. Critically, this classifier was then tested with supertrials from the remaining six animate and six inanimate exemplars in their recognizable form. These procedures ensured that the number of trials used for training and testing were exactly the same as those used for decoding the same stimulus category, thus are similarly powered. We also conducted cross-decoding in the opposite direction (training on recognizable originals, testing on texforms).

To create a graphical depiction of the similarity structure in the measured EEG responses, we used the following approach. First, electrode patterns were extracted for each object exemplar at each time point, yielding 60 conditions for recognizable images and 60 conditions for texform images. Next, we measured the multivariate noise-normalized Euclidean distance (Guggenmos, Sterzer, & Cichy, 2018) between EEG patterns of all possible object pairs. Therefore, a 120 × 120 representational dissimilarity matrix was obtained for each participant at each time point. Finally, we used multidimensional scaling (MDS) to transform the group-averaged EEG representational

dissimilarity matrix at the peak decoding time into a 2-D space. Note that these plots are purely a supplementary visualization to provide a graphical intuition of the successful cross-decoding results (e.g., the main decoding analyses were not conducted in this 2-D MDS space).

Pairwise Decoding

To determine the decodability of each object against others, we estimated the pairwise decoding performance of all pairs of objects for both original and texform images. Linear discriminant analysis classifiers were trained and evaluated with ROC AUC metric via five iterations of cross-validation. On each iteration, we trained a classifier to discriminate between two objects on 80% of trials and tested on the held-out 20% of trials. Please note that no supertrial averaging was applied here because of the limited number of trials for each single object stimulus (original: 22.6 ± 1.4; texform: 22.9 ± 1.3). The final pairwise decoding performance at each time point was the average of all pairwise decoding results across all cross-validation attempts (1770 pairs × 5 iterations). For the sake of saving computation time, we downsampled the EEG data with a decimation factor of two.

Statistical Testing

To examine whether the decoding performance was significantly above chance, we conducted one-sided Wilcoxon signed-rank tests, which is nonparametric and does not make any assumptions about the shape of the data distribution. When comparing the performance of different conditions of interest, we used two-sided Wilcoxon signed-rank tests. We conducted these statistical tests across the time points in a time window of interest (100–500 msec) and then applied false discovery rate (FDR) correction ($p < .05$). The time window of interest was determined as the duration of a 400-msec presentation with a starting point at 100 msec when the stimuli have full onset (stimuli were faded-in in the first 100 msec).

The latency of decoding onset was defined as the first time point with above-chance decoding ($p < .01$, uncorrected) for three consecutive time points; this approach was adapted from several previous studies (Robinson, Grootswagers, & Carlson, 2019; Cichy et al., 2014; Carlson et al., 2013). Note that in this procedure, multiple comparisons are not applied so that the estimation of onset latency does not depend on the decoding performance of later time points. The time of peak decoding was defined as the time point with maximum performance within the time window of interest (100–500 msec). In the case where there were multiple local maximums within the window, the first of those maximums was selected.

We assessed the median and confidence interval of the onset and peak latencies using bootstrap sampling (with replacement) with 5000 iterations (for a similar analysis, see Robinson et al., 2019; Cichy, Pantazis, & Oliva, 2016; Cichy et al., 2014). To test the differences of onset and peak latencies, we estimated the p values based on bootstrapped distributions. Such results were corrected for the number of comparisons using FDR correction with the significance level of $p < .05$.

RESULTS

Animate versus Inanimate Decoding

First, we examined whether recognizable images of animate and inanimate objects evoked distinguishable spatial EEG patterns over time, as has been previously shown (e.g., Khaligh-Razavi et al., 2018; Grootswagers, Wardle, et al., 2017; Ritchie, Tovar, & Carlson, 2015; Carlson et al., 2013). Figure 1B (solid silver line) shows a plot of decoding accuracy as a function of time for original images. Consistent with previous work, we observed a robust ability to classify animacy information: The spatial topography of the elicited EEG responses to animate and inanimate recognizable images were distinguishable from each other (ps $< .05$, one-sided signed-rank test, FDR corrected), with significant onset at 126 msec (95% CI [116, 142] msec) and peak classification accuracy at 188 msec (95% CI [184, 200] msec).

Next, we investigated (i) whether unrecognizable texform images of animate and inanimate objects evoke distinct spatial EEG patterns, and if so, (ii) at what time these distinctions emerge relative to the recognizable image counterparts. The same classification analysis as above was performed but considering only responses to texform images (Figure 1B, solid black line). Animate and inanimate texforms elicited different EEG patterns, with an onset of significant decoding at 152 msec (95% CI [106, 164] msec) and an early classification peak at 176 msec (95% CI [146, 190] msec). The onset and peak latencies of decoding for texform images did not significantly differ from those for recognizable images (onset: $p = .34$, peak: $p = .14$, bootstrapping test, FDR corrected). Critically, animacy decoding did not emerge over several hundreds of milliseconds, as would be predicted if extra processing time was needed to extract and/or amplify animacy information from texform images. However, animacy decoding did have a lower accuracy for texforms in comparison to original images (non-independent peak decoding accuracy: original 74.46% vs. texform 56.01%, $p < .001$, two-sided signed-rank test). Overall, these results indicate that the mid-level feature content preserved in texform images contains early perceptual signatures of animacy information.

Are the features that support the animacy distinction in texforms the same as those supporting animacy decoding in original recognizable images? If this is the case, both texforms and original images should evoke the same topographical differences that distinguish between animate and inanimate objects. To test this possibility, we conducted cross-decoding analyses in which we trained classifiers to discriminate EEG responses to animate versus

inanimate texform images, and then tested the classifiers on responses to animate and inanimate original images. To ensure that classifiers were generalizing to new examples, we did not include any of the original-counterpart images to the texforms used to train the classifier. As shown in Figure 1B (dashed gray line), we found that texform-trained classifiers could successfully classify whether a new recognizable object was animate or inanimate (ps < .05, one-sided signed-rank test, FDR corrected). Such successful decoding was also evident early (onset: 140 msec, 95% CI [114, 152] msec; peak: 176 msec, 95% CI [174, 192] msec), with no significant difference in time to original images (onset: p = .44, peak: p = .34, bootstrapping test, FDR corrected) or texform images (onset: p = .14, peak: p = .46, bootstrapping testing, FDR corrected). Moreover, we also observed similar results when conducting the cross-decoding in the opposite direction (training on recognizable originals and testing on texforms). Thus, the classification boundary between animate and inanimate texforms also separates the animate and inanimate recognizable images, demonstrating the activation patterns are similar between these image formats.

Unexpectedly, we found that texform-trained classifiers could predict the animacy more accurately for recognizable images than for other texform images (non-independent peak decoding: texform-original 63.4% vs. texform-texform 56.0%, p < .001, two-tailed signed-rank test). How is this superior classification accuracy possible? One possibility is that the original images evoke more discriminable neural responses than texforms, while still sharing a common large-scale topographic decision boundary. Consistent with this possibility, Figure 1B (right) provides a graphical intuition for this explanation. This MDS plot visualizes the neural pattern similarity structure among the original images (open dots) and texform images (filled dots) at the peak cross-decoding time (176 msec), such that items with similar neural response patterns are nearby in the plot. Note that there is a general separation between animates (purple dots) and inanimates (green dots) across both texforms and originals. Furthermore, the texform images (filled dots) are closer to each other; in contrast, recognizable images (open dots) are more distinctive and farther apart in this visualization. Thus, this visualization helps provide an intuition for how original images can be classified more accurately than texform images by a texform-trained classifier.

A second piece of evidence also supports the interpretation that the original images evoke more separable, distinctive neural responses than those evoked by texforms. Specifically, we estimated the discriminability of responses at the item level, estimating the average pairwise decoding accuracy over all pairs of items. Figure 2 shows that pairwise decoding accuracy is significantly higher for original images than for texform images (ps < .05, two-sided signed-rank test, FDR corrected). Thus, we reason that, to the degree that both texforms and original images evoke similar patterns of neural responses that share a common decision boundary, the original images should be more easily classifiable because of their more distinctive evoked brain responses. In this way, this cross-decoding result provides strong evidence that the differences of elicited spatial topography that reflect animacy distinction in texforms are highly compatible with the distinguishing differences between recognizable animals and objects.

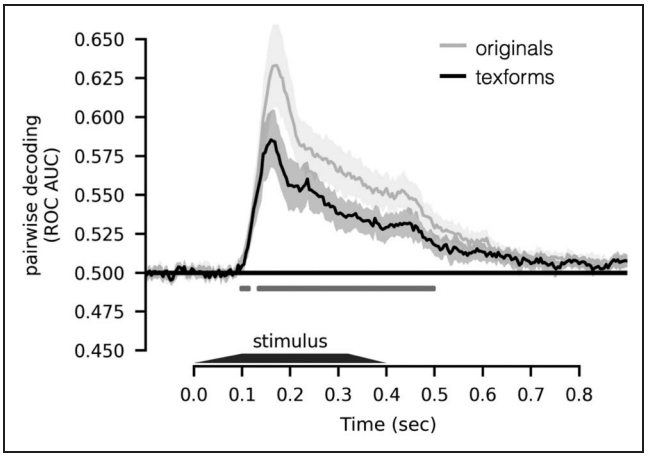

Figure 2. Time course of pairwise decoding. Pairwise decoding performance averaged across all object pairs is plotted along the y axis, as a function of time (x axis), for originals (silver line) and texforms (black line); shaded region indicates 95% CI. Time points with significant difference between original and texform stimuli are depicted below the time courses (two-sided signed-rank test, ps < .05, FDR corrected in the time window of interest, 100–500 msec).

Big versus Small Decoding

Next, we examined evoked differences between big and small entities. Overall, the results reveal a similar pattern of results but with weaker overall decoding accuracy, plotted in Figure 1C (left). There was a significant difference between the elicited EEG response patterns to original images depicting big entities and small entities (ps < .05, one-sided signed-rank test, FDR corrected), as well as for texform images (ps < .05, one-sided signed-rank test, FDR corrected). The timing of this emerging size distinction was also early in the response: neither decoding onsets nor decoding peaks for texform responses (onset: 130 msec, 95% CI [120, 246] msec; peak: 150 msec, 95% CI [114, 162] msec) were significantly different from those for original images (onset: 120 msec, 95% CI [114, 132] msec, p = .38; peak: 174 msec, 95% CI [110, 194] msec, p = .88, bootstrapping test, FDR corrected), although we note the lower accuracy is also accompanied with less confident estimates of the onset. Furthermore, we found significant cross-decoding evident in classifiers trained on texform images and tested on original images (ps < .05, one-sided signed-rank test, FDR corrected), also evident early in time (onset: 124 msec, 95% CI [120, 128] msec;

peak: 140 msec, 95% CI [130, 172] msec), with no significant difference in timing to original images (onset: p = .38, peak: p = .88, bootstrapping test, FDR corrected) or to texform images (onset: p = .38; peak: p = .88, bootstrapping test, FDR corrected). In summary, the above results demonstrate systematic (albeit weak) differences in neural responses to texformed versions of big and small images, evident early in the time course of processing, with compatible EEG response structure as evoked by original images.

We next conducted further analysis to assess size decoding separately for the animate and inanimate domains, motivated by previous work with fMRI by Konkle and Caramazza (2013). In particular, the spatial activations of ventral visual cortex exhibit three large-scale cortical zones preferentially responding to big inanimate objects, small inanimate objects, and animals (of both sizes). That is, there were similar spatial activation patterns for big and small animals (Konkle & Caramazza, 2013). Thus, we next examined the degree to which this "tripartite" signature was also apparent in the decoding of EEG responses. Given these previous findings from fMRI, we expected size decoding to be stronger among inanimate objects than among animals.

The results are shown in Figure 3 (top). We found that size information was decodable from responses evoked by inanimate objects, and by animate objects, for both originals and texforms (all ps < .05, one-sided signed-rank test, FDR corrected). However, size decoding from responses to animal images was actually stronger than size decoding from responses to object images, contrary to what we expected (ps < .05, two-sided signed-rank test, FDR corrected). Note that this pattern of results held in both texforms and originals images. Considering the time course of this size decoding, responses to big versus small animals show an earlier and more rapid rise in their classifiability, whereas responses to big versus small inanimate objects show a slower and more gradual separability.

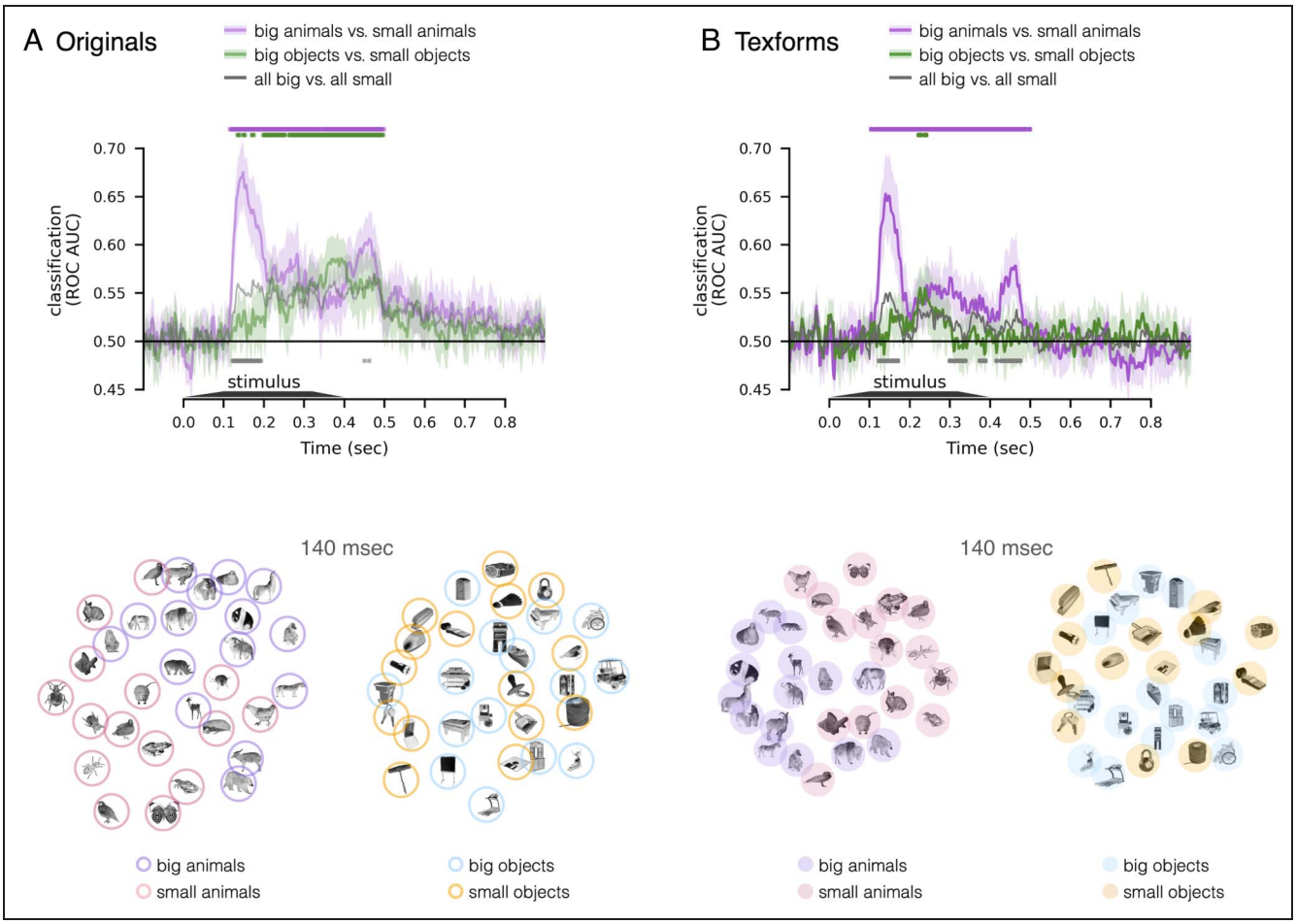

Figure 3. Decoding size among animate or inanimate objects only, for (A) original images and (B) texform images. In both plots (upper), classification accuracy (y axis) is plotted as a function of time (x axis). Purple line: big animals versus small animals. Green line: big objects versus small objects. Gray line: the combined classification for animates and inanimates is plotted for reference, which corresponds to the silver (original) or black (texform) line in Figure 1C. Time points with significant decoding are depicted above the time courses (ps < .05, one-sided signed-rank test, FDR corrected in the time window of interest, 100–500 msec), and time points with significant difference between original and texform stimuli are depicted below the time courses (ps < .05, two-sided signed-rank test, FDR corrected). Below the line plots are MDS visualizations, with a 2-D projection of the pairwise distances of the neural responses to animate objects only or inanimate objects only, examined at the peak cross-decoding (140 msec).

Using MDS, we visualized the EEG pattern similarity structure among animals and among inanimate objects, separately for original and texform images (Figure 3 bottom). Specifically, we visualized the similarity structure evident at 140 msec, when the texform-to-original cross-decoding showed peak performance. In line with the decoding results, this visualization shows that the separation between big and small objects are clearer for animate objects in comparison to inanimate objects (for both original and texform images). Thus, these EEG size decoding results reveal a notable difference between the scalp-electrode response patterns over time and the large-scale cortical activation patterns along the ventral pathway that aggregated over time. We speculate on the underlying causes of these patterns of data in the Discussion section.

DISCUSSION

Here, we employed multivariate EEG decoding to examine whether and when the visual system is sensitive to mid-level feature differences related to the broad distinctions of animacy and real-world size. We used a well-established stimulus set that includes recognizable images of big and small animals and objects, as well as their unrecognizable "texform" counterparts (Long et al., 2016, 2017, 2018). We found that: (1) neural responses measured by EEG to texform images contained early information about animacy and size, as evident by above-chance decoding accuracy. (2) This broad categorical information was decodable from evoked responses to texforms at a similar time as from evoked responses to recognizable original images. (3) In addition, the time-evolving activation patterns were similar between these image formats, as evident by significant cross-decoding, suggesting a common underlying basis. Broadly, these EEG results indicate that the visual system contains an extensive mid-level feature bank, with early sensitivity to mid-level feature differences supporting animacy and size distinctions.

These patterns of data, and our subsequent interpretations, offer a different perspective than recent work by Grootswagers, Robinson, Shatek, et al. (2019). Specifically, Grootswagers, Robinson, Shatek, et al. (2019) also explored if animacy and size could be decoded from texform images, but they employed a fast image presentation paradigm (Grootswagers, Robinson, & Carlson, 2019) in which the presentation rate was varied from 5 Hz to 60 Hz—differing from our slow event-related design. In their data, texforms elicited brain response structure with weaker decoding of animacy information than recognizable objects, and only at the slowest presentation rate. Based on these results, they proposed that additional processing time in higher order visual areas is required to further "untangle" the mid-level feature differences evident in texforms into linearly separable categorical organizations (cf. DiCarlo & Cox, 2007). In contrast, we propose that no further "untangling" is required for animacy and object size information to emerge.

To reconcile our findings with Grootswagers, Robinson, Shatek, et al. (2019), we offer the following possibility. We propose that the visual system contains a mid-level feature bank that carries linearly decodable information on animacy and size. Texforms and original images rapidly activate this feature bank in a primarily feedforward processing sweep, enabling early decoding. However, perhaps when stimuli are presented in rapid succession without gap time in between, as in Grootswagers, Robinson, Shatek, et al. (2019), the recurrent/feedback activity from the previous stimulus interferes with the early processing stages of the incoming stimulus. For example, back-to-back presentations have been reported to elicit smaller periodic signals (Retter, Jiang, Webster, & Rossion, 2018) and delayed neural responses (Robinson et al., 2019) in comparison to presentation schedules with gap time between successive stimuli. We also observed in our data that texforms do not elicit the same magnitude of feature activation as original images—this is evident in our data by their generally lower decoding accuracy, both at the category and item-level, and is also found in neuroimaging results (Long et al., 2018). Thus, these responses may be more likely to be extinguished under conditions of forward masking, leading to accentuated differences between original and texform images. Thus, rather than requiring more untangling time, our proposal accounts for the similarities between texforms and originals seen in our study at early time points, and instead posits increased susceptibility to forward masking during rapid texform presentation.

One other pattern of these data was that animacy decoding was more accurate from neural responses to original stimuli than to texform stimuli—what factors might underlie this accuracy difference? One possibility is that the original stimuli have additional mid-level visual features not captured by the texform generation algorithm (e.g., clear outer and inner contours). It is important to keep in mind that texforms preserve some mid-level visual features related to second-order image statistics in localized pooling regions, but these are not necessarily a perfect model of mid-level visual representation. Relatedly, another possibility is that decoding was higher for original images because they contain additional category-specific object parts that are not present in texforms. For example, animals often have tails, eyes, and noses, and these object parts are obscured in the texform images. Finally, participant attention may have differed between these two sets of stimuli, as recognizable original stimuli may better capture attention than texform stimuli. These possibilities are not mutually exclusive. Further studies are needed to determine what stimulus properties and task effects account for the animacy decoding gap between original and texform stimuli.

How do the current real-world size decoding results relate to previous fMRI work? Specifically, Konkle and Caramazza (2013) found that big and small object images evoked a large-scale organization of responses across the

cortical surface, whereas big versus small animals had similar response topographies. We expected that EEG decoding accuracy would also reflect this tripartite organization, but that is not what we found. We can rule out the possibility that the distinction between big and small animals was driven by the detection of recognizable eyes or frontal faces, because this result was also evident in the texform images, which lack clear facial features. One possibility, invited by the time-course of decoding, is that the neural populations that distinguish between big and small animals are only engaged early and transiently, and their responses may not be evident in slower aggregated responses of fMRI. This spatial–temporal hypothesis may be possible to explore through fMRI-magnetoencephalography fusion (Khaligh-Razavi et al., 2018; Cichy et al., 2016), electrocorticography, or neural recordings in monkey populations. More generally, these results highlight the need for a deeper exploration of the convergences and discrepancies between the spatial similarity structure of neural activation patterns over EEG electrodes, and BOLD-estimated activations over cortical voxels.

Although texform and original stimuli both quickly evoked neural responses with information on animacy and size, one limitation of this study relates to the precision at which we could measure the onset latencies of these decoding results. In some cases, the onset latency of decoding had a 95% confidence interval spanning several tens of milliseconds, making it difficult to detect subtle differences in the timing of decoding results. Such ranges of variability could arise from individual differences in decoding time course and have also been observed in other studies that have reported the confidence intervals of onset latency (Robinson et al., 2019; Cichy et al., 2014, 2016). Because of this variability, we would interpret the onset time with some level of caution. In this study, the timing of animacy and size decoding for both original and texform stimuli is compatible with an early, primarily feedforward stage of processing, rather than protracted recurrent processing evolving over hundreds of milliseconds. However, subtle differences between texforms and original images on the order of tens of milliseconds may not have been revealed by our methods. Another limitation is that the number of stimuli employed was relatively limited ($n = 60$, 15 per animacy-size combination), leaving open the possibility that these randomly selected exemplars may not be fully representative of the broader categories they were sampled from. In our analyses, we leveraged cross-validation methods that require predicting animacy and size in held-out stimuli, mitigating this concern with an analytical approach.

This work joins a growing set of results showing the tight links between original and texformed counterparts in perceptual processes(e.g., Chen, Deza, & Konkle, 2022; Long et al., 2016, 2017, 2018) and more generally between mid-level feature distinctions and broader categorical distinction (Groen, Silson, & Baker, 2017). Overall, this work provides clear support for the claim that early visual processes operating over mid-level features contain information about the broad categorical distinctions of animacy and object size.

Acknowledgments

We thank Aylin Kallmayer and Hrag Pailian for their help during the experiments and data collection. This work was supported by NSF CAREER BCS-1942438 (T. K.).

Reprint requests should be sent to Ruosi Wang or Daniel Janini, Department of Psychology, Harvard University, 33 Kirkland st. 7 floor, Cambridge, MA, 02138, United States, or via e-mail: wang.ruosi@outlook.com or janinidp@gmail.com.

Author Contributions

R. W., D. J., and T. K. designed research. R. W. and D. J. performed research. R. W. analyzed data. R. W., D. J., and T. K. interpreted the results. R. W. and T. K., wrote the first draft of the paper. R. W., D. J., and T. K. edited the paper. R. W. and D. J. contributed unpublished analytic tools.

Funding Information

Talia Konkle, Division of Behavioral and Cognitive Sciences (https://dx.doi.org/10.13039/100000169), grant number: CAREER BCS-1942438.

Diversity in Citation Practices

Retrospective analysis of the citations in every article published in this journal from 2010 to 2021 reveals a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the *Journal of Cognitive Neuroscience* (*JoCN*) during this period were M(an)/M = .407, W(oman)/M = .32, M/W = .115, and W/W = .159, the comparable proportions for the articles that these authorship teams cited were M/M = .549, W/M = .257, M/W = .109, and W/W = .085 (Postle and Fulvio, *JoCN*, 34:1, pp. 1–3). Consequently, *JoCN* encourages all authors to consider gender balance explicitly when selecting which articles to cite and gives them the opportunity to report their article's gender citation balance.

Notes

1. In the initial stimulus set from Long et al. (2018), there were 120 texforms total, 30 for each of the animacy × size conditions. These 30 images were further split into six groups based on their level of classifiability, reflecting how well independent participants could guess whether the texform was animate/inanimate and big/small. We used a subset of these stimuli by randomly selected stimuli from each level of classifiability: three exemplars from each group with highest, high, and medium–high classifiability; and two exemplars from each group with medium–low, low, and lowest classifiability.
2. Early in piloting, we tested this paradigm both with our 64-channel EEG system and with a custom channel configuration

with more electrodes over the visual cortex. These equipment changes did not yield any differences in the overall pattern of our pilot data. Thus, we went to the 32-channel system because the setup time was much shorter, which enabled us to increase the power per subject within the limited duration of an EEG experimental session.

REFERENCES

Bae, G.-Y., & Luck, S. J. (2018). Dissociable decoding of spatial attention and working memory from EEG oscillations and sustained potentials. *Journal of Neuroscience*, *38*, 409–422. https://doi.org/10.1523/JNEUROSCI.2860-17.2017, PubMed: 29167407

Baldassi, C., Alemi-Neissi, A., Pagan, M., DiCarlo, J. J., Zecchina, R., & Zoccolan, D. (2013). Shape similarity, better than semantic membership, accounts for the structure of visual object representations in a population of monkey inferotemporal neurons. *PLoS Computational Biology*, *9*, e1003167. https://doi.org/10.1371/journal.pcbi.1003167, PubMed: 23950700

Bao, P., She, L., McGill, M., & Tsao, D. Y. (2020). A map of object space in primate inferotemporal cortex. *Nature*, *583*, 103–108. https://doi.org/10.1038/s41586-020-2350-5, PubMed: 32494012

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436. https://doi.org/10.1163/156856897X00357, PubMed: 9176952

Carlson, T., Tovar, D. A., Alink, A., & Kriegeskorte, N. (2013). Representational dynamics of object vision: The first 1000 ms. *Journal of Vision*, *13*, 1. https://doi.org/10.1167/13.10.1, PubMed: 23908380

Cauchoix, M., Crouzet, S. M., Fize, D., & Serre, T. (2016). Fast ventral stream neural activity enables rapid visual categorization. *Neuroimage*, *125*, 280–290. https://doi.org/10.1016/j.neuroimage.2015.10.012, PubMed: 26477655

Chen, Y.-C., Deza, A., & Konkle, T. (2022). How big should this object be? Perceptual influences on viewing-size preferences. *Cognition*, *225*, 105114. https://doi.org/10.1016/j.cognition.2022.105114, PubMed: 35381479

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, *17*, 455–462. https://doi.org/10.1038/nn.3635, PubMed: 24464044

Cichy, R. M., Pantazis, D., & Oliva, A. (2016). Similarity-based fusion of MEG and fMRI reveals spatio-temporal dynamics in human cortex during visual object recognition. *Cerebral Cortex*, *26*, 3563–3579. https://doi.org/10.1093/cercor/bhw135, PubMed: 27235099

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*, 333–341. https://doi.org/10.1016/j.tics.2007.06.010, PubMed: 17631409

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience*, *14*, 1195–1201. https://doi.org/10.1038/nn.2889, PubMed: 21841776

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., et al. (2014). MNE software for processing MEG and EEG data. *Neuroimage*, *86*, 446–460. https://doi.org/10.1016/j.neuroimage.2013.10.027, PubMed: 24161808

Grill-Spector, K., & Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, *15*, 536–548. https://doi.org/10.1038/nrn3747, PubMed: 24962370

Groen, I. I. A., Silson, E. H., & Baker, C. I. (2017). Contributions of low- and high-level properties to neural processing of visual scenes in the human brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*, 20160102. https://doi.org/10.1098/rstb.2016.0102, PubMed: 28044013

Grootswagers, T., Ritchie, J. B., Wardle, S. G., Heathcote, A., & Carlson, T. A. (2017). Asymmetric compression of representational space for object animacy categorization under degraded viewing conditions. *Journal of Cognitive Neuroscience*, *29*, 1995–2010. https://doi.org/10.1162/jocn_a_01177, PubMed: 28820673

Grootswagers, T., Robinson, A. K., & Carlson, T. A. (2019). The representational dynamics of visual objects in rapid serial visual processing streams. *Neuroimage*, *188*, 668–679. https://doi.org/10.1016/j.neuroimage.2018.12.046, PubMed: 30593903

Grootswagers, T., Robinson, A. K., Shatek, S. M., & Carlson, T. A. (2019). Untangling featural and conceptual object representations. *Neuroimage*, *202*, 116083. https://doi.org/10.1016/j.neuroimage.2019.116083, PubMed: 31400529

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2017). Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data. *Journal of Cognitive Neuroscience*, *29*, 677–697. https://doi.org/10.1162/jocn_a_01068, PubMed: 27779910

Guggenmos, M., Sterzer, P., & Cichy, R. M. (2018). Multivariate pattern analysis for MEG: A comparison of dissimilarity measures. *Neuroimage*, *173*, 434–447. https://doi.org/10.1016/j.neuroimage.2018.02.044, PubMed: 29499313

Isik, L., Meyers, E. M., Leibo, J. Z., & Poggio, T. (2014). The dynamics of invariant object recognition in the human visual system. *Journal of Neurophysiology*, *111*, 91–102. https://doi.org/10.1152/jn.00394.2013, PubMed: 24089402

Jagadeesh, A. V., & Gardner, J. L. (2022). Texture-like representation of objects in human visual cortex. *Proceedings of the National Academy of Sciences*, *119*, e2115302119. https://doi.org/10.1073/pnas.2115302119, PubMed: 35439063

Jas, M., Engemann, D. A., Bekhti, Y., Raimondo, F., & Gramfort, A. (2017). Autoreject: Automated artifact rejection for MEG and EEG data. *Neuroimage*, *159*, 417–429. https://doi.org/10.1016/j.neuroimage.2017.06.030, PubMed: 28645840

Jozwik, K. M., Kriegeskorte, N., & Mur, M. (2016). Visual features as stepping stones toward semantics: Explaining object similarity in IT and perception with non-negative least squares. *Neuropsychologia*, *83*, 201–226. https://doi.org/10.1016/j.neuropsychologia.2015.10.023, PubMed: 26493748

Julian, J. B., Ryan, J., & Epstein, R. A. (2017). Coding of object size and object category in human visual cortex. *Cerebral Cortex*, *27*, 3095–3109. https://doi.org/10.1093/cercor/bhw150, PubMed: 27252351

Kaneshiro, B., Perreau Guimaraes, M., Kim, H.-S., Norcia, A. M., & Suppes, P. (2015). A representational similarity analysis of the dynamics of object processing using single-trial EEG classification. *PLoS One*, *10*, e0135697. https://doi.org/10.1371/journal.pone.0135697, PubMed: 26295970

Khaligh-Razavi, S.-M., Cichy, R. M., Pantazis, D., & Oliva, A. (2018). Tracking the spatiotemporal neural dynamics of real-world object size and animacy in the human brain. *Journal of Cognitive Neuroscience*, *30*, 1559–1576. https://doi.org/10.1162/jocn_a_01290, PubMed: 29877767

Konkle, T., & Caramazza, A. (2013). Tripartite Organization of the Ventral Stream by Animacy and object size. *Journal of Neuroscience*, *33*, 10235–10242. https://doi.org/10.1523/JNEUROSCI.0983-13.2013, PubMed: 23785139

Konkle, T., & Oliva, A. (2012). A real-world size organization of object responses in occipitotemporal cortex. *Neuron*, *74*, 1114–1124. https://doi.org/10.1016/j.neuron.2012.04.036, PubMed: 22726840

Long, B., Konkle, T., Cohen, M. A., & Alvarez, G. A. (2016). Mid-level perceptual features distinguish objects of different real-world sizes. *Journal of Experimental Psychology: General*, *145*, 95–109. https://doi.org/10.1037/xge0000130, PubMed: 26709591

Long, B., Störmer, V. S., & Alvarez, G. A. (2017). Mid-level perceptual features contain early cues to animacy. *Journal of Vision*, *17*, 20. https://doi.org/10.1167/17.6.20, PubMed: 28654965

Long, B., Yu, C.-P., & Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences, U.S.A.*, *115*, E9015–E9024. https://doi.org/10.1073/pnas.1719616115, PubMed: 30171168

Mahon, B. Z., Anzellotti, S., Schwarzbach, J., Zampini, M., & Caramazza, A. (2009). Category-specific organization in the human brain does not require visual experience. *Neuron*, *63*, 397–405. https://doi.org/10.1016/j.neuron.2009.07.012, PubMed: 19679078

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peelen, M. V., & Downing, P. E. (2017). Category selectivity in human visual cortex: Beyond visual object recognition. *Neuropsychologia*, *105*, 177–183. https://doi.org/10.1016/j.neuropsychologia.2017.03.033, PubMed: 28377161

Proklova, D., Kaiser, D., & Peelen, M. V. (2016). Disentangling representations of object shape and object category in human visual cortex: The animate–inanimate distinction. *Journal of Cognitive Neuroscience*, *28*, 680–692. https://doi.org/10.1162/jocn_a_00924, PubMed: 26765944

Retter, T. L., Jiang, F., Webster, M. A., & Rossion, B. (2018). Dissociable effects of inter-stimulus interval and presentation duration on rapid face categorization. *Vision Research*, *145*, 11–20. https://doi.org/10.1016/j.visres.2018.02.009, PubMed: 29581059

Ritchie, J. B., Tovar, D. A., & Carlson, T. A. (2015). Emerging object representations in the visual system predict reaction times for categorization. *PLoS Computational Biology*, *11*, e1004316. https://doi.org/10.1371/journal.pcbi.1004316, PubMed: 26107634

Ritchie, J. B., Zeman, A. A., Bosmans, J., Sun, S., Verhaegen, K., & Op de Beeck, H. P. (2021). Untangling the animacy organization of occipitotemporal cortex. *Journal of Neuroscience*, *41*, 7103–7119. https://doi.org/10.1523/JNEUROSCI.2628-20.2021, PubMed: 34230104

Robinson, A. K., Grootswagers, T., & Carlson, T. A. (2019). The influence of image masking on object representations during rapid serial visual presentation. *Neuroimage*, *197*, 224–231. https://doi.org/10.1016/j.neuroimage.2019.04.050, PubMed: 31009746

Stormer, V. S., Alvarez, G. A., & Cavanagh, P. (2014). Within-Hemifield competition in early visual areas limits the ability to track multiple objects with attention. *Journal of Neuroscience*, *34*, 11526–11533. https://doi.org/10.1523/JNEUROSCI.0980-14.2014, PubMed: 25164651

Thorat, S., Proklova, D., & Peelen, M. V. (2019). The nature of the animacy organization in human ventral temporal cortex. *eLife*, *8*, e47142. https://doi.org/10.7554/eLife.47142, PubMed: 31496518

Vinken, K., Konkle, T., & Livingstone, M. (2022). The neural code for 'face cells' is not face specific. *bioRxiv*. https://doi.org/10.1101/2022.03.06.483186

Yue, X., Robert, S., & Ungerleider, L. G. (2020). Curvature processing in human visual cortical areas. *Neuroimage*, *222*, 117295. https://doi.org/10.1016/j.neuroimage.2020.117295, PubMed: 32835823