Bottom-up processing of curvilinear visual features is sufficient for animate/inanimate object categorization

Valentinos Zachariou
Laboratory of Brain and Cognition, NIMH/NIH, Bethesda, MD, USA

Amanda C. Del Giacco
Laboratory of Brain and Cognition, NIMH/NIH, Bethesda, MD, USA

Leslie G. Ungerleider
Laboratory of Brain and Cognition, NIMH/NIH, Bethesda, MD, USA

Xiaomin Yue
Laboratory of Brain and Cognition, NIMH/NIH, Bethesda, MD, USA

Animate and inanimate objects differ in their intermediate visual features. For instance, animate objects tend to be more curvilinear compared to inanimate objects (e.g., Levin, Takarae, Miner, & Keil, 2001). Recently, it has been demonstrated that these differences in the intermediate visual features of animate and inanimate objects are sufficient for categorization: Human participants viewing synthesized images of animate and inanimate objects that differ largely in the amount of these visual features classify objects as animate/inanimate significantly above chance (Long, Stormer, & Alvarez, 2017). A remaining question, however, is whether the observed categorization is a consequence of top-down cognitive strategies (e.g., rectangular shapes are less likely to be animals) or a consequence of bottom-up processing of their intermediate visual features, per se, in the absence of top-down cognitive strategies. To address this issue, we repeated the classification experiment of Long et al. (2017) but, unlike Long et al. (2017), matched the synthesized images, on average, in the amount of image-based and perceived curvilinear and rectilinear information. Additionally, in our synthesized images, global shape information was not preserved, and the images appeared as texture patterns. These changes prevented participants from using top-down cognitive strategies to perform the task. During the experiment, participants were presented with these synthesized, texture-like animate and inanimate images and, on each trial, were required to classify them as either animate or inanimate with no feedback given. Participants were told that these synthesized images depicted abstract art patterns. We found that participants still classified the synthesized stimuli significantly above chance even though they were unaware of their classification performance. For both object categories, participants depended more on the curvilinear and less on the rectilinear, image-based information present in the stimuli for classification. Surprisingly, the stimuli most consistently classified as animate were the most dangerous animals in our sample of images. We conclude that bottom-up processing of intermediate features present in the visual input is sufficient for animate/inanimate object categorization and that these features may convey information associated with the affective content of the visual stimuli.

Introduction

Human and nonhuman primates are remarkably fast at visual object categorization (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976; Perrett, Hietanen, Oram, Benson, & Rolls, 1992; Thorpe, Fize, & Marlot, 1996; Hung, Kreiman, Poggio, & DiCarlo, 2005; Grill-Spector & Kanwisher, 2005; Cauchoix, Crouzet, Fize, & Serre, 2016). Given the number of processing stages involved and the speed at which object categorization occurs, this visual process is believed to be predominantly feed-forward (Riesenhuber & Poggio, 1999; Serre, Oliva, & Poggio, 2007). That is, visual information is processed hierarchically at increasing levels of abstraction beginning with edge extraction in the primary visual cortex (V1, e.g., Hubel & Wiesel, 1959) to the processing of intermediate visual features (e.g., by area V4; Gallant, Braun, & Van Essen, 1993; Yue, Pourladian, Tootell, & Ungerleider, 2014) to the

Citation: Zachariou, V., Del Giacco, A. C., Ungerleider, L. G., & Yue, X. (2018). Bottom-up processing of curvilinear visual features is sufficient for animate/inanimate object categorization. *Journal of Vision*, *18*(12):3, 1–12, https://doi.org/10.1167/18.12.3.

https://doi.org/10.1167/18.12.3 Received April 23, 2018; published November 6, 2018 ISSN 1534-7362

processing of more complex visual features (Tanaka, 1996) and/or object categories (e.g., Tsao, Freiwald, Tootell, & Livingston, 2006; Kriegeskorte et al., 2008; Bell, Hadj-Bouziane, Frihauf, Tootell, & Ungerleider, 2009; Bi, Wang, & Caramazza, 2016) in the inferior temporal (IT) cortex, where different regions respond selectively to different object categories (Haxby et al., 2001; Kanwisher, 2010). More recently, however, it has been proposed that basic object categorization, such as distinguishing between animate and inanimate objects, may not depend exclusively on processing by regions in the IT cortex; instead, processing by intermediate visual areas, which compute intermediate object features, might be sufficient for this distinction (Perrinet & Bednar, 2015; Long, Konkle, Cohen, Alvarez, 2016; Long et al., 2017). This hypothesis advocates that animate and inanimate object categories differ substantially in their intermediate visual features, and these differences are sufficient for animate/inanimate categorization. For instance, animate objects are more curved compared to inanimate objects (Levin et al., 2001; Perrinet & Bednar, 2015; Long et al., 2017). As such, the amount of curvilinear information present in the visual input may carry adequate information for human and nonhuman primates to distinguish between animate and inanimate objects with minimal IT cortex involvement. Evidence supporting this hypothesis comes from a recent study (Long et al., 2017) demonstrating that participants could visually classify animate and inanimate objects significantly above chance based on the amount of curvilinear and rectilinear information present in the images: The authors generated synthetic stimuli (termed *texforms*) from images of animals (animate objects) and man-made objects (inanimate objects) using a texture synthesis algorithm described in detail in Freeman and Simonelli (2011). This algorithm (model) extracts a group of image statistical descriptors across spatial scales and orientations, including mean luminance, contrast, skewness, and kurtosis, from spatially constrained windows throughout a target image. The algorithm then iteratively adjusts the pixel values of a random Gaussian noise image with the same spatial dimensions as the original image using a variant of gradient descent until this noise image has the same image statistical descriptors within the same spatially constrained windows as the original image. By pooling and synthesizing image statistics from and within separate but spatially constrained windows, the resulting stimuli preserve only the coarse form of the target object but maintain the object's intermediate features (e.g., curvilinear, rectilinear, and some texture information). These texform images could not be recognized as the original object but carried sufficient information to be classified above chance as animate and inanimate.

Even though the findings of Long et al. (2017) are intriguing, it is unclear whether the participants used top-down cognitive strategies or bottom-up visual processing to perform the classification task. Although the texform stimuli used in that study could not be recognized as the original objects, a substantial amount of global shape information was preserved in the images. For instance, many of the stimuli in the inanimate category had texforms with rectangular shapes, and in fact, the participants perceptually rated the texform images in the inanimate category as more "boxy" and those in the animate category as more "curvy" (see Figure 1). Thus, participants may have used a simple, top-down cognitive strategy to classify the images (e.g., rectangular shapes are less likely to be animals) given that the experimental instructions stated that the texform stimuli were created by "scrambling" animate or man-made objects.

Consequently, a remaining question is whether bottom-up processing of intermediate visual features, per se, in the absence of top-down cognitive strategies is sufficient for animate/inanimate categorization. Here, we examined this possibility using a procedure very similar to the animate/inanimate classification task of Long et al. (2017) with two important modifications. First, we matched the animate and inanimate synthesized stimuli, on average, on the perceived and computationally calculated amount of curvilinear and rectilinear information. Consequently, participants could not rely on overall differences in the amount of these features between the two object categories to perform the task. Second, we eliminated global shape information from the synthesized images (see Figures 1 and 2), preventing participants from using coarse shape information (e.g., circular, rectangular, etc.) to perform the task. Under these conditions, above-chance classification was only possible when the curvilinear/rectilinear features present in each individual synthesized image conveyed sufficient information for animate/inanimate categorization.

Our findings are mostly consistent with, but greatly extend, those of Long et al. (2017): We found that, under the matched conditions of our experiment, participants could still classify the synthesized animate/inanimate stimuli significantly above chance. Importantly, however, the participants' confidence ratings of their classification performance did not predict classification accuracy. Therefore, it is unlikely that participants used top-down cognitive strategies to perform the tasks: Participants were not aware of their classification performance. In contrast to Long et al. (2017), classification accuracy was only predicted by the amount of the image-based (calculated; see Methods) curvilinear information present in each image, indicating that the curvilinear information present in the visual input was more important for

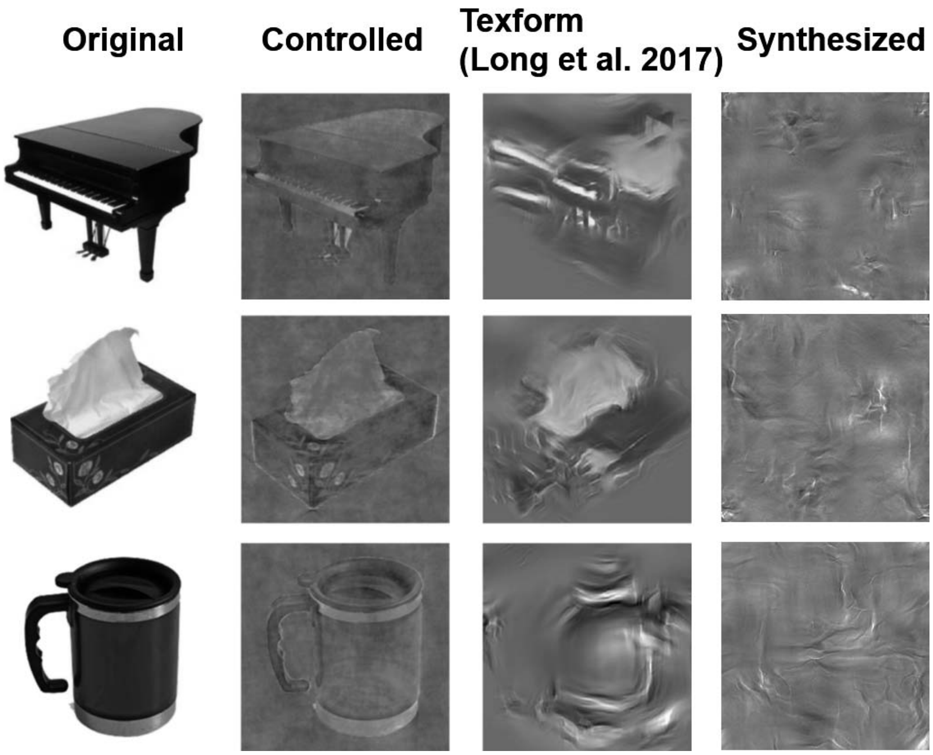

Figure 1. Three example inanimate objects used in Long et al. (2017) together with (a) their corresponding texform images, created using the Freeman and Simoncelli (2011) algorithm, (b) their corresponding synthesized images created using the Portilla and Simoncelli (2000) algorithm. The images under the "original," "controlled," and "texform" columns were directly extracted from figure 1 of Long et al. (2017). The images under the "texform" column were created using the algorithm described in Freeman and Simoncelli (2011) with some slight modifications outlined in Long et al. (2016). The images under the "synthesized" column were created by using the algorithm described in Portilla and Simoncelli (2000), which we used in this study. Both the "texform" and "synthesized" object algorithms used the images under the "controlled" column as inputs.

classification compared to the rectilinear information. Finally and unexpectedly, the synthesized animate images with higher classification accuracy were the more dangerous animals in our sample of images, implying that some affect-related information was contained in these intermediate visual features.

Materials and methods

Experiment design

The experiment consisted of three behavioral sessions: two image-rating sessions and an image-classification session. During the rating sessions, two separate groups of participants ($n = 15$ per group) rated the synthesized animate and inanimate images: One group of participants rated the images on the amount of curvilinearity (how curvy each image was), and the other group rated the images on amount of rectilinearity (how boxy/rectangular each image was). During the classification session, a different group of participants ($n = 20$), who did not participate in the first two sessions, classified the synthesized images as either animate or inanimate.

Participants

Fifty healthy adults were recruited for the experiment (33 females, age range 21–34 years). Thirty of these (19 females, age range 22–34 years) participated in the two rating sessions, and the remaining 20 (14 females, age range 21–28 years) participated in the classification session. All participants were right-handed with normal or corrected vision. All gave informed consent under a protocol approved by the institutional review board of the National Institute of Mental Health.

Visual stimuli

First, 105 animate and 178 inanimate images were downloaded from the Internet (using various search terms on Google Images). The animate images com-

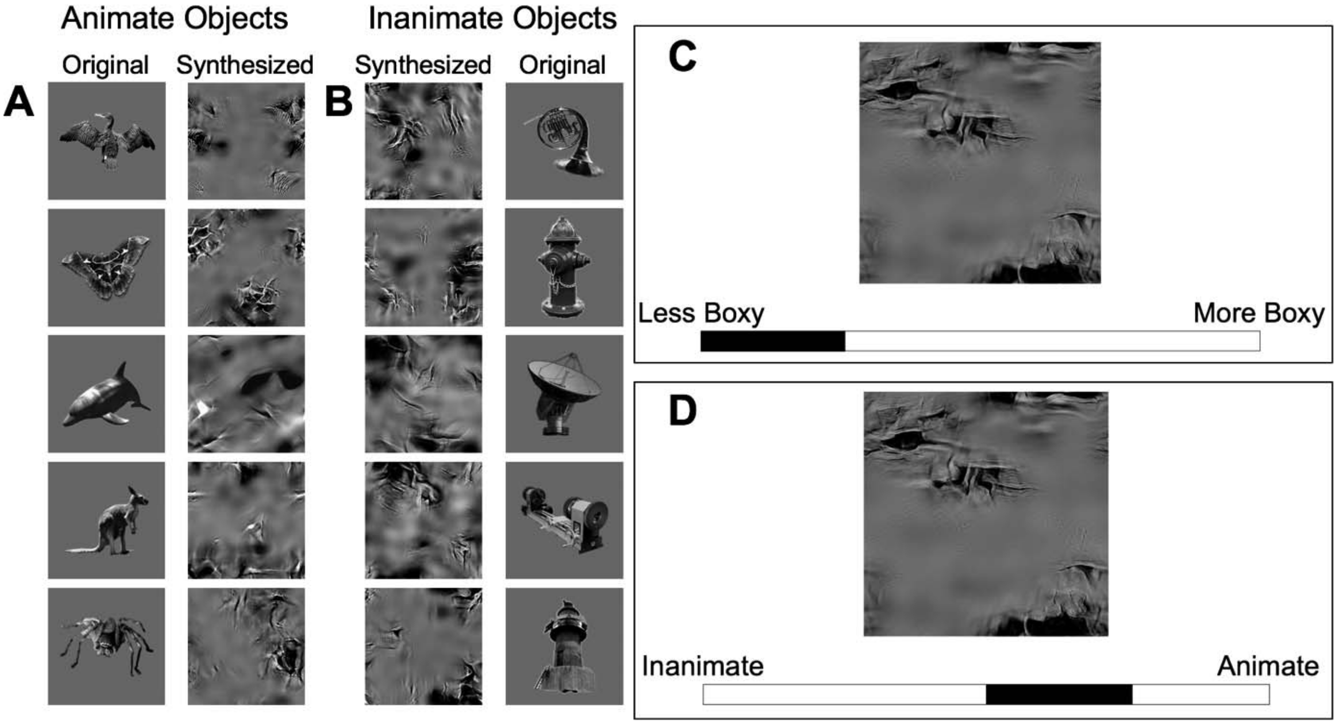

Figure 2. Example stimuli and sample trial displays from the rating and classification sessions of the experiment. (A) Examples of animate images with their corresponding synthesized stimuli. (B) Examples of inanimate/man-made images with their corresponding synthesized stimuli. (C) A sample trial display from the rating session of the experiment. This example depicts a trial in which a participant rated the synthesized stimuli on degree of "boxiness" or how rectilinear the images appeared to him or her. The black bar corresponds to the amount of "boxiness" this participant attributed to the stimulus image. (D) Sample trial display from the classification session of the experiment. This example depicts a trial in which a participant classified the synthesized stimulus as "animate." The black bar represents the confidence of the participant on his or her classification choice.

prised mammals, birds, fish, reptiles, and insects (Figure 2A). The inanimate images comprised man-made objects, such as tools, vehicles, buildings, and various household items (Figure 2B). All object images were digitally processed (see Supplementary Material for a detailed description of this process) to have the same size, background, mean luminance, and root-mean-square contrast.

Quantifying the amount of curvilinear and rectilinear information of the stimuli

After matching the stimuli on size, background, mean luminance, and contrast, we calculated the amount of curvilinear and rectilinear information present in each image using a method very similar to the one presented previously in Yue et al. (2014; a detailed description of this procedure is given in the Supplementary Material).

Using these calculated values of curvilinear and rectilinear information, we selected a smaller set of images from the entire sample of 105 animate and 178 inanimate objects with comparable amounts of rectilinear and curvilinear information. This smaller set consisted of 29 images per object category for a total of 58 images.

Creating synthesized images

We used an algorithm described in detail in Portilla and Simoncelli (2000), to generate synthesized images (Figure 2A and B) of animate and inanimate objects, which maintained the intermediate visual features of the original images. This algorithm uses steerable pyramid filters to extract a group of image statistical descriptors, including first-, second-, third-, and fourth-order image statistics, from a target input image at the level of the whole image in a spatially unconstrained manner. These image statistics include mean luminance, contrast, skewness, and kurtosis. The algorithm then permutes pixel values of a random Gaussian noise image with the same dimensions as the original image across multiple scales and orientations (using a variant of gradient descent) until the image statistical descriptors of the noise image converge to those of the original image.

The main difference between the algorithm used here and the Freeman and Simoncelli (2011) algorithm used in Long et al. (2017) is as follows: In the algorithm we used (Portilla & Simoncelli, 2000), the image statistical descriptors between an original image and its corresponding synthesized image are matched in the global image dimension. Conversely, the Freeman and Simoncelli algorithm matches these image statistical descriptors at the level of receptive field–like, spatially constrained windows (similar in size to those of human area V2). This is the main reason why the texform stimuli in Long et al. (2017) preserve the coarse form of the original objects.

Power spectrum analysis

We conducted a power spectrum analysis to quantify the contrast energy carried by the high, median, and low spatial frequency bands of each of the synthesized images. This analysis allowed us to examine (see Results) whether spatial frequency predicted classification accuracy. The detailed description of this analysis is given in the Supplementary Material.

Experimental procedure

The experiment, implemented using E-prime 2.0, was run on a Windows 7–based PC. The stimuli were presented on a desktop monitor (BENQ XL) at a resolution of 1,024 × 768 pixels (0.022° per pixel). The viewable area of the display was 407 × 307 mm (23° visual angle horizontally by 17° vertically at a distance of 100 cm away from the participants' eyes).

Each rating session lasted 5 min. The presentation order of the rating sessions, i.e., whether a participant rated the synthesized images on amount of curvilinear or rectilinear information, was counterbalanced across participants. During these sessions, each of the 58 images (29/category) was presented once at the center of the screen in randomized order against a white background at a resolution of 384 × 384 pixels (159 × 159 mm; 9° × 9° visual angle) for 10 s. Directly below each image on the display was a horizontal rectangular bar, which was initially white at the onset of each trial (matching the background color; Figure 2C). During the trial, participants filled this bar with black by moving a mouse to the right. Moving the mouse to the left decreased the portion of black that filled the bar. Depending on the rating session (curvilinear or rectilinear ratings), participants were instructed that the portion of the rectangular bar filled with black either represented the amount of rectilinear (how box-like an image looked) or curvilinear information present in an image (how curvy an image looked). Specifically, for the curvilinear rating group, participants were instructed to evaluate how many simple and/or complex curve-like features (we showed participants examples of these during the instructions) made up an image. The rectilinear rating group was instructed to evaluate how many triangular and/or rectangular features (examples of these were also shown during the instructions) made up an image. As such, the area of the bar covered with black indicated a participant's perceptual rating of how curvilinear or rectilinear they thought each image was. Participants left-clicked the mouse to record their response, and immediately afterward, the next trial/image was presented. Participants were also instructed to respond as quickly as possible and that if they failed to respond within 10 s from the onset of a trial that trial aborted and their response was marked as "wrong." This last instruction was provided in order to encourage participants to respond within the time limit. The area of the rectangular bar filled with black on each trial was converted to a percentage of the total area of the rectangle and recorded for later analysis.

During the classification session, performed by a different group of participants, the synthesized stimuli were presented at the same on-screen location with the same white background and at the same size as in the rating sessions. Each of the 58 images was presented eight times (464 trials) in randomized order with a 1-min break after the first four presentations of each image (after the first 232 trials). In four of these image presentations (randomly selected), the synthesized images appeared upright, and in the remaining four, the images appeared mirror-reversed and upside down to prevent as much as possible the participants from memorizing the stimuli. Each stimulus was presented for 5 s. If a participant did not respond within the time limit, that trial aborted and was marked as a no-response trial. Like the rating sessions, participants responded by selecting how much of a rectangular bar, situated right below the images on the display, would be filled with black (Figure 2D). For the classification session, however, the rectangular bar was split into two equal portions between the left- and right-hand sides of the vertical center of the display. At stimulus onset, both the left and right sides of the rectangular bar were white. Participants could increase the amount of black on each side of the rectangular bar by moving the mouse either to the left or to the right. The black area always increased away from the center of the display and toward either the left or the right side of the display. For instance, moving the mouse to the right gradually increased the amount of the right part of the rectangle that filled with black; moving the mouse to the left gradually decreased the amount of the right part of the rectangle filled with black until the vertical center of the display. If the mouse continued to move in the same direction, the left portion of the rectangular bar would gradually fill with black. For half the participants, the right side of the rectangular bar represented the animate category and the left side of the bar represented the inanimate category. For the other half of the participants, the order was reversed. On each trial, a text label, located on both ends of the rectangular response bar, indicated which side of the bar corresponded to the animate and which side to the inanimate category. The amount of the left or right side of the rectangular bar filled with black indicated each participant's confidence in his or her response. For example, if a participant wished to classify a stimulus as animate on a given trial, the participant was required to move the mouse in the direction that filled the side of the bar corresponding to the animate category. The area of the bar (left or right side) filled represented his or her confidence on how animate the stimulus image appeared. For each trial, we recorded a participant's binary choice (animate or inanimate) depending on which side of the bar a participant decided to fill with black and a confidence rating associated with his or her choice. Participants were given no feedback during the experiment regarding their classification accuracy.

The difference in the stimulus duration between the curvilinearity/rectilinearity rating sessions (10 s) and the classification task session (5 s) is because, on average, participants were faster to classify the images as animate/inanimate (average response time: 1,933 ms, $SD = 942$ ms) than to rate the images on curvilinearity/rectilinearity (average response time: 3,509.5 ms, $SD =$ 1,567.5 ms). The reason participants were slower during the rating sessions is because they were instructed to evaluate how many simple and/or complex curvilinear- or rectilinear-like features made up each image, which forced them to visually search each image for these features. To this end, we selected a longer stimulus duration for the rating sessions to allow ample time for the participants to complete the task.

Statistical analyses

The behavioral measures collected during the rating and classification sessions were analyzed as follows. Excel was used to calculate *d*-prime values and criterion *C* values for each participant. Independent, paired samples and one-sample *t* tests (two-tailed) as well as linear mixed effects ANOVAs were used in SPSS to analyze the participants' classification accuracy (percentage correct), *d*-prime values, criterion *C* values, and confidence ratings. Bayesian *t* tests were also used in SPSS to analyze the data when necessary. Linear regressions were also conducted in SPSS in order to evaluate whether or not the curvilinear and rectilinear information present in the images and/or the participants' confidence ratings predicted classification performance. Multiple comparisons used Sidak corrections when necessary.

Additionally, we conducted two different types of permutation tests on the classification accuracy of the participants in MATLAB (MathWorks, Natick, MA). One set of permutation tests was performed at the individual participant level and another at the group level. The permutation tests at the individual participant level allowed us to account for the response bias of each participant separately and to subsequently examine how many of the participants classified above, below, or at chance level. The group-level permutation tests allowed us to examine classification accuracy separately for each object category (animate/inanimate) by accounting for the response bias of the participants at the group level. A detailed description of these permutation tests can be found in the Supplementary Material.

Results

Matching the amount of image-based curvilinear and rectilinear information between the two object categories

A series of control analyses contrasting the amount of image-based and perceived curvilinear and rectilinear information between the two object categories can be found in the Supplementary Material. As anticipated, these analyses did not reveal significant differences between the two object categories: average curvilinearity of the animate images: 0.94, $SD = 0.3$, $SEM = 0.012$; of the inanimate images: 0.90, $SD = 0.33$, $SEM = 0.012$; average rectilinearity of the animate images: 0.76, $SD = 0.24$, $SEM = 0.008$; of the inanimate images: 0.82, $SD = 0.27$, $SEM = 0.009$.

It should be noted that there was a significant positive correlation between the perceived and image-based measures of curvilinearity, $t(1, 57) = 1.92$, $p =$ 0.03, $r = 0.28$. This correlation, however, explained only 8% of the variance ($r^2 = 0.078$). The correlation between the perceived and image-based measures of rectilinearity was not significant, $t(1, 57) = -0.01$, $p = 0.993$, $r =$ 0.001. An internal reliability analysis on the perceived measures of curvilinearity (Cronbach's alpha = 0.68) and rectilinearity (Cronbach's alpha = 0.73) revealed that the reliability of these perceived measures was relatively low (Tavakol & Dennick, 2011). For this internal reliability analysis, the perceived ratings of each image were averaged separately for even- and odd-numbered participants and used in the calculation of Cronbach's alpha. In summary, there appears to be a weak relationship between the perceived and image-based measures of curvilinearity/rectilinearity and, as expected, the image-based (calculated) measures are more sensitive and more accurate than the corresponding perceived measures.

Classification accuracy

Here we evaluated the classification accuracy of the participants in categorizing the synthesized images as animate or inanimate. We used three different tests to explore the participants' classification performance: (a) For each participant, we calculated his or her overall percentage correct accuracy. We then compared these percentage correct scores to chance performance (50% correct) using a one-sample *t* test in SPSS. (b) Using the animate category as "signal," we calculated *d*-prime and criterion *C* values for each participant. We then contrasted these *d*-prime values against "zero" using a one-sample *t* test in SPSS. (c) Last, we conducted permutation tests at both the individual participant and group levels in order to evaluate if participants' classification was above chance irrespective of object category or whether above chance classification was for only one of the two object categories. Additionally, we used the permutation tests at the individual participant level to evaluate how many of the participants' classifications were above chance, at chance, and below chance (consistently misclassified).

The one sample *t* test (two-tailed) contrasting the participants' overall classification accuracy (percentage correct) to chance (50%) was significant, such that the average accuracy for classification (54.62% accurate) was significantly above chance, $t(1, 19) = 3.931$, $p = 0.001$, $SEM = 1.1\%$, Bayes factor = 40.21, prior 0.7071. Similarly, the one-sample *t* test (two-tailed) contrasting the participants' *d*-prime scores to zero (chance performance) was also significant: The average *d*-prime score of the participants (0.22) was significantly greater than zero, $t(1, 19) = 3.395$, $p = 0.003$, $SEM = 0.065$, Bayes factor = 14.02, prior 0.7071. The criterion *C* of the participants ($C = 0.153$) approached significance but was not significantly different from zero, $t(1, 19) = 1.915$, $p = 0.071$, $SEM = 0.08$, Bayes factor = 1.06, prior 0.7071.

Participants were 12% more likely to classify an image as inanimate versus animate. As such, we performed permutation tests (see Methods) to account for this response bias. The group-level permutation tests (see Supplementary Material for a detailed description) indicated that the participants' overall accuracy (54.62%) occurred by chance with a probability $p = 0.008$. The classification accuracy for the animate category occurred by chance with a probability $p < 0.0001$ (accuracy adjusted for the participants' bias using the individual participant–level permutation tests = 54.83%). The classification accuracy for the inanimate object category occurred by chance with a similar probability of $p < 0.0001$ (accuracy adjusted for the participants bias using the individual participant–level permutation tests = 52.78%). A paired samples *t* test (two-tailed) on classification accuracy, adjusted for the response bias of each object category, indicated that the two did not differ significantly, $t(1, 19) = 0.328$, $p = 0.747$, $SEM = 6.25\%$, Bayes Factor = 0.249, prior = 0.7071.

Further, using the permutation tests at the individual-subject level, we binned participants into those who classified the object categories significantly above chance, at chance, and significantly below chance. This distribution was as follows: 13 participants classified the images above chance, four classified the images at chance, and three participants classified the images significantly below chance.

Last, in order to evaluate the internal reliability of the classification accuracy of the participants, we performed the following analysis. First, we split the participants into two groups based on whether their participant number was even or odd. Then, for each of these even/odd groups, we calculated the classification accuracy for each of the 58 synthesized images in the animate and inanimate object categories. We then used these per-image classification accuracies obtained from the even- and odd-numbered participant groups to calculate Cronbach's alpha. The overall Cronbach's alpha, across all synthesized images for both object categories was 0.87. The Cronbach's alpha for the synthesized animate images only was 0.83 and for the inanimate images was 0.91. To this end, classification accuracy was reliable across participants (Tavakol & Dennick, 2011).

In summary, the participants' classification accuracy and *d*-prime scores were significantly above chance: 65% of the participants classified the images significantly above chance, 20% classified the images at chance, and the remaining 15% of the participants misclassified the images. At the group level, participants classified both object categories significantly above chance, and their classification performance did not differ significantly between the two object categories.

The contribution of top-down cognitive strategies to classification accuracy

To evaluate whether or not the participants were consciously aware of their classification performance, which would imply that they used top-down cognitive strategies to perform the task, we conducted a linear mixed effects ANOVA with the participants' classification accuracy, confidence ratings during classification, and object category (animate or inanimate) as factors. The confidence ratings for each trial were recorded as a continuous variable (between 0% and 100% for each category). We used a histogram to bin the confidence scores into five bins in order to calculate the classification accuracy for each bin, separately for each object category and for each participant, and used these values in the ANOVA. We split the confidence ratings into five bins in order to have a comparable number of trials across bins. This ANOVA did not yield any significant main effects or an interaction between the factors: object category, $F(1, 172) = 0.747$, $p = 0.561$; confidence rating bin, $F(1, 172) = 0.682$, $p = 0.605$; object category × confidence rating bin, $F(1, 172) = 0.581$, $p = 0.677$. Consequently, classification accuracy did not change significantly across confidence rating bins, irrespective of object category.

As an additional step to evaluate whether or not the participants' confidence predicted their accuracy, we correlated the participants' overall classification accuracy with their overall level of confidence. Across participants, this correlation was also not significant, $t(1,19) = 0.562$, $p = 0.581$, $r = 0.13$.

Last, we conducted a linear regression ANCOVA between the overall classification accuracy of each image and the corresponding confidence ratings of the images with object category (animate or inanimate) as an interaction term. Object category, however, did not interact significantly with the correlation between the classification accuracy of each image and the corresponding confidence rating ($p = 0.507$). Similarly, the correlation between the classification accuracy of each image and the corresponding confidence rating was not significant, $t(1, 57) = -0.813$, $p = 0.419$, $r = 0.11$.

In summary, the participants' confidence ratings did not predict their classification accuracy. Additionally, the classification accuracy of each image was not predicted by its corresponding confidence rating. Thus, there is no evidence that the participants were consciously aware of their classification performance.

The contribution of image-based curvilinear and rectilinear information to classification accuracy

We next attempted to identify what information the participants used to perform the classification task. Here, using a linear regression ANCOVA, we evaluated the correlation between the participants' classification accuracy for each synthesized image and the amount of image-based curvilinear and rectilinear information present in the image (see Methods). Further, we explored if/how this correlation varied between the two object categories. For this analysis, we excluded the perceived measures of curvilinearity and rectilinearity because, as described earlier, the internal reliability of these perceived measures was relatively low.

We found a significant interaction of the correlation between the classification accuracy of each image and the amount of image-based curvilinear information present in the image across object category ($p < 0.0001$). To unpack this significant interaction, we conducted regression analyses separately for each object category (Figure 3). For the animate category, the amount of image-based curvilinear information in the images positively predicted their classification accuracy, $t(1, 28) = 2.32$, $p = 0.028$, $r_{\text{partial (control for rectilinear)}} = 0.408$; the more curvilinear the animate image, the more accurately it was classified. For the inanimate category, the amount of image-based curvilinear information in the images also predicted classification accuracy, but this correlation was negative, $t(1, 28) = -3.05$, $p = 0.005$, $r_{\text{partial (control for rectilinear)}} = -0.506$; the less curvilinear an inanimate image, the more accurately it was classified.

The analysis on the image-based rectilinear information, however, did not reveal a significant interaction of the correlation between the classification accuracy of each image and the amount of image-based rectilinear information present in the image across object category ($p = 0.106$). Additionally, the amount of image-based rectilinear information present in the images did not predict classification accuracy, $t(1, 56) = 6.56$, $p = 0.335$, $r = 0.13$.

Taken together, our data show that the amount of image-based curvilinear information present in the images, but not the amount of rectilinear information, predicted the classification accuracy for both animate and inanimate object categories.

The contribution of spatial frequency to classification accuracy

We also explored whether participants used spatial frequency information to classify the two object categories. For this analysis, we conducted separate linear regression ANCOVAs for each object category between the classification accuracy for each image and the power carried by (a) the low spatial frequencies in each image, (b) the median spatial frequencies in each image, and (c) the high spatial frequencies in each image. None of these linear regression analyses, however, yielded significant results—animate object stimuli: low spatial frequency power, average power = 1.16, $SD = 0.25$, $t(1, 28) = 0.205$, $p = 0.839$, $r = 0.04$; median spatial frequency power, average power = 0.92, $SD = 0.42$, $t(1, 28) = 0.547$, $p = 0.589$, $r = 0.105$; high spatial frequency power, average power = 0.66, $SD = 0.44$, $t(1, 28) = -0.218$, $p = 0.829$, $r = -0.042$; inanimate object stimuli: low spatial frequency power, average power = 0.31, $SD = 0.09$, $t(1, 28) = -0.820$, $p = 0.419$, $r = 0.156$; median spatial frequency power, average power = 0.20, $SD = 0.13$, $t(1, 28) = -0.189$, $p = 0.851$, r

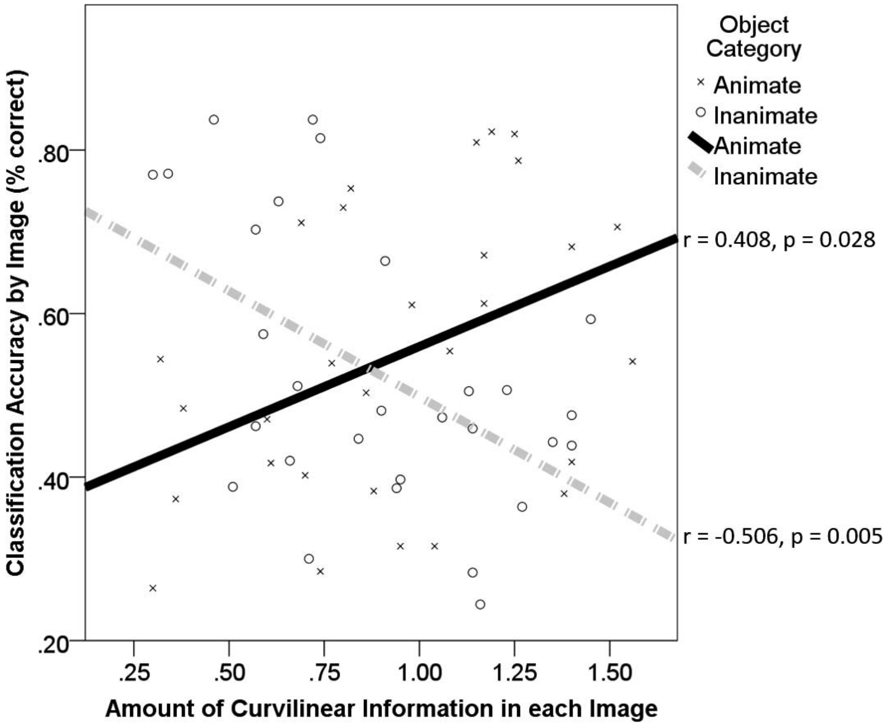

Figure 3. The figure depicts how the classification accuracy of the animate (X) and inanimate (O) images varied as a function of the amount of calculated curvilinear information present in the images. The best-fit line for the animate category is represented by the solid black line, and the best-fit line for the inanimate category is represented by the dotted gray line. The average classification accuracy was 54.62%, which is significantly above chance.

$= 0.036$; high spatial frequency power, average power $=$ 0.12, $SD = 0.09$, $t(1, 28) = -0.228$, $p = 0.821$, $r = 0.044$.

In summary, the spatial frequency of the synthesized stimuli did not predict the participants' classification accuracy, making it unlikely that participants used this information for classification.

The relationship between the affective content of the images and classification accuracy

By rank-ordering the synthesized animate images according to classification accuracy (Figure 4), we found that participants classified the dangerous animals (e.g., a scorpion, a tarantula, a shark) more accurately compared to less dangerous animals (e.g., a dolphin, a butterfly, a kangaroo). To more carefully examine this unexpected finding, we assigned valence and arousal ratings to each animate image (separate ratings for valance and arousal per image) using the International Affective Picture System (IAPS; Lang, Bradley, & Cuthbert, 2008). A detailed description of how valance and arousal scores were obtained for the IAPS database images can be found in Lang et al. (2008). For this assignment, we used the ratings from the "all subjects" section of the IAPS database. Six out of the 29 animate images in our sample did not have a corresponding image in the IAPS database and were excluded. When one of the synthesized animate images in our sample corresponded to more than one image in the IAPS database, that image was assigned the average valence and arousal ratings of all the corresponding images in IAPS. For example, one of our animate images depicted a shark, but the IAPS database has three different shark images (and one shark image depicted together with a diver, which was excluded), and so our shark image was assigned the average valance and arousal level of the three shark images in the IAPS database. Then, using a linear regression analysis, we correlated the classification accuracy of the synthesized animate images with their assigned valance and arousal scores.

The valence scores of the synthesized animate images correlated negatively with classification accuracy, $t(1, 28) = -2.330$, $p = 0.028$, $r_{\text{partial (controlling for arousal)}} =$ -0.422; the lower the valence (more unpleasant), the higher the classification accuracy. The arousal ratings of the animate images did not correlate with classification accuracy, $t(1, 28) = 0.771$, $p = 0.448$, $r_{\text{(controlling for valance)}} = 0.152$. To evaluate whether the significant correlation we observed between classification accuracy and valance was due to visual differences (as measured in this study) between the animate stimuli, we repeated this analysis within the context of a partial correlation, controlling for image-based curvilinearity;

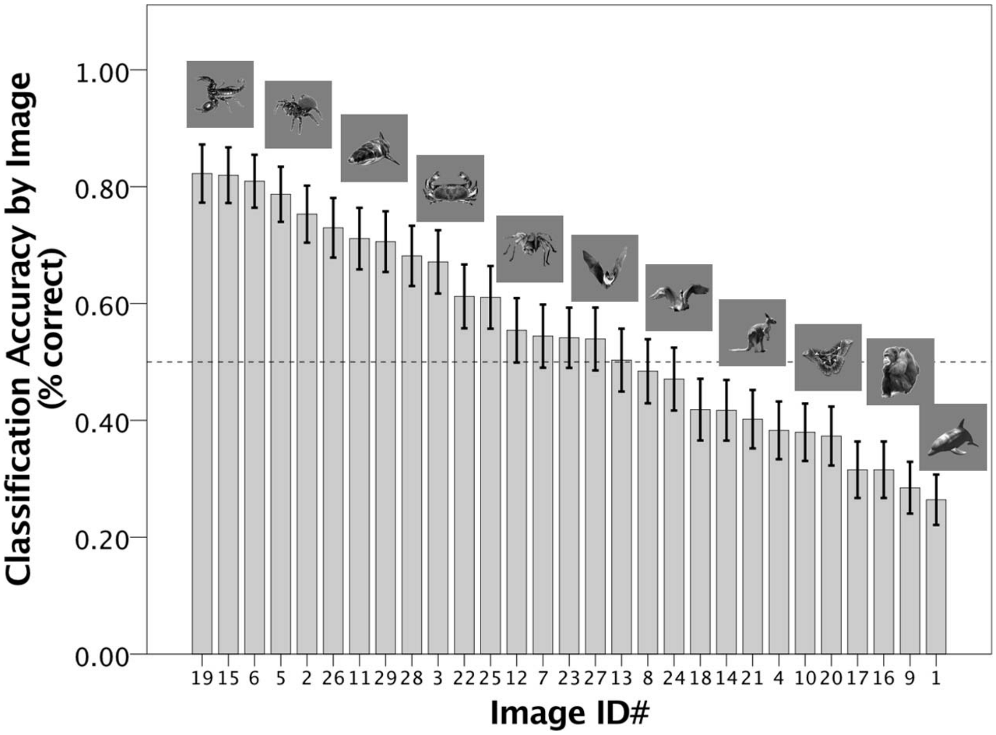

Figure 4. Rank-ordered classification accuracy for the synthesized animate images in descending order. A sample of the original images, corresponding to the rank-ordered image IDs of the synthesized images, are presented on top of the bars. The horizontal dashed line represents chance performance (50% accuracy). The error bars denote ±1 *SEM*.

rectilinearity; and low, median, and high spatial frequencies. The correlation between classification accuracy and valance, however, was not eliminated when we controlled for these factors, $t(1, 28) = -2.1$, $p = 0.04$, $r_{\text{partial}} = -0.41$.

In summary, this significant correlation between the classification accuracy and the valance ratings of the images implies that intermediate visual features may convey information about the affective content/threat level of animals in addition to object category. This possibility will be explored in greater detail in a future study.

Discussion

The current experiment was designed to address whether bottom-up processing of intermediate visual features in the absence of top-down cognitive strategies is sufficient for animate and inanimate object categorization. We found that (a) despite the two object categories being matched on average on the amount of perceived and image-based curvilinear and rectilinear information, participants still classified the synthesized stimuli significantly above chance, and (b) confidence ratings during classification did not predict classification accuracy. Therefore, participants were not consciously aware of their classification performance, making it unlikely that they used top-down cognitive strategies to perform this task. (c) The curvilinear but not the rectilinear information present in the images predicted classification performance; and (d) the animate stimuli that were classified more accurately and more consistently as animate were synthesized from the dangerous animal images (i.e., those with lower valence IAPS ratings), whereas those animate stimuli with lower classification accuracy were synthesized from images of less dangerous animals (those with higher valence IAPS ratings). Because we did not find any evidence of top-down cognitive strategies, it appears that the above-chance classification of the synthesized animate and inanimate stimuli depended primarily on bottom-up visual processing of their curvilinear features. Presumably this processing was mediated by intermediate visual areas in the ventral visual pathway (e.g., by V4) as no coherent shape information was present in the visual input.

It should be noted that, in accord with our findings, Long et al. (2017) also concluded, based on a set of EEG data, that their participants relied on early visual/bottom-up processing rather than top-down cognitive strategies to complete the visual tasks. One concern, however, with this conclusion is that the EEG data described in Long et al. (2017) were collected during a visual search task and not during the classification task.

In this EEG search task, participants had to visually search for an animate or inanimate target texform among a varying number of distracters of the same animacy (animate or inanimate texform distracters) or mixed-animacy distracters. As illustrated in Figure 1, global shape information was not completely eliminated in the texform stimuli of Long et al. (2017) due to the properties of the algorithm they used. Importantly, the inanimate texform stimuli of Long et al. (2017) were rated by participants as more "boxy" and the animate texform stimuli as more "curvy." As such, it remains possible that the visual search task can be expressed in the form of searching for rectangular-/curvilinear-shaped targets among rectangular/curvilinear distracters or searching for rectangular-/curvilinear-shaped targets among mixed displays consisting of rectangular and curvilinear distracters. Under these conditions, unlike the classification task, top-down cognitive strategies are not required, and the visual search task could be performed based on bottom-up visual processing of the difference in global shape between the animate and inanimate texform targets and distracters.

Our findings are consistent with previous studies suggesting that processing of intermediate visual features influences high-level object categorization (e.g., Bar & Neta, 2006; Cheung & Gauthier, 2014; Perrinet & Bednar, 2015; Cauchoix et al., 2016; Long & Konkle, 2017; Schmidt, Hegele, & Fleming, 2017). For example, Perrinet and Bednar (2015), using data from the rapid masked, animal versus nonanimal categorization task of Serre et al. (2007), found that the images most consistently miscategorized as containing an animal consisted of second-order image statistics very similar to those images that contained animals and very different image statistics from those images that did not.

One possible hypothesis to explain why intermediate visual features might contain early cues to animacy is that of survival. Almost all animals, both predators and prey, exhibit some sort of natural camouflage, which allows them to blend in with their surroundings. From an evolutionary perspective, being able to rapidly determine with little visual information if something hidden in the environment is alive and/or dangerous would benefit the survival of both predators (better prey detection) and prey (better predator detection). Our findings are consistent with this idea. Participants consistently classified the synthesized stimuli corresponding to the more dangerous animals as more animate. In the future, it will be interesting to examine whether the synthesized stimuli corresponding to the dangerous animals within the animate category activate the amygdala or other affect-processing systems of the brain more strongly compared to the synthesized stimuli of the less-threatening animals.

Conclusion

In conclusion, we found that bottom-up processing of the curvilinear features, per se, in the absence of top-down cognitive strategies, conveys sufficient information for animate/inanimate object categorization. In contrast to the findings of Long et al. (2017), the rectilinear information present in the visual input did not predict classification accuracy, suggesting that participants based their categorization predominantly on the curvilinear rather than the rectilinear visual information for both the animate and inanimate object categories. Last, the intermediate visual features of the animate object category appear to convey information associated with the valance of the stimuli. This latter, unexpected finding will be explored in more detail in future studies.

Keywords: object categorization, *curvilinear features*, *rectilinear features*, *valance*, *arousal*

Acknowledgments

We thank Merage Ghane for assistance with the IAPS database, and Drs. Alex Martin and Zvi Roth for thoughtful comments on the manuscript. This research was supported by the NIMH Intramural Research Program (ZIA MH002918-09) under the NIH Clinical Study Protocol 93-M-0170 (NCT00001360).

Commercial relationships: none.

Corresponding author: Valentinos Zachariou; Xiaomin Yue.
E-mail: vzachari@gmail.com; yuex@mail.nih.gov.
Address: Laboratory of Brain and Cognition, NIMH/NIH, Bethesda, MD, USA.

References

Bar, M., & Neta, M. (2006). Humans prefer curved visual objects. *Psychological Science*, *17*(8), 645–648.

Bell, A. H., Hadj-Bouziane, F., Frihauf, J. B., Tootell, R. B., & Ungerleider, L. G. (2009). Object representations in the temporal cortex of monkeys and humans as revealed by functional magnetic resonance imaging. *Journal of Neurophysiology*, *101*(2), 688–700.

Bi, Y., Wang, X., & Caramazza, A. (2016). Object domain and modality in the ventral visual pathway. *Trends in Cognitive Sciences*, *20*(4), 282–290.

Cauchoix, M., Crouzet, S. M., Fize, D., & Serre, T. (2016). Fast ventral stream neural activity enables rapid visual categorization. *NeuroImage*, *125*, 280–290.

Cheung, O. S., & Gauthier, I. (2014). Visual appearance interacts with conceptual knowledge in object recognition. *Frontiers in Psychology*, *5*, 793.

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience*, *14*(9), 1195–1201.

Gallant, J. L., Braun, J., & Van Essen, D. C. (1993, January 1) Selectivity for polar, hyperbolic, and Cartesian gratings in macaque visual cortex. *Science*, *259*, 100–103

Grill-Spector, K., & Kanwisher, N. (2005). Visual recognition: As soon as you know it is there, you know what it is. *Psychological Science*, *16*(2), 152–160.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001, September 28). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–2430.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, *148*(3), 574–591.

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005, November 4). Fast readout of object identity from macaque inferior temporal cortex. *Science*, *310*(5749), 863–866.

Kanwisher, N. (2010). Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences, USA*, *107*(25), 11163–11170.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., . . . Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–1141.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8. Gainesville, FL: University of Florida.

Levin, D. T., Takarae, Y., Miner, A. G., & Keil, F. (2001). Efficient visual search by category: Specifying the features that mark the difference between artifacts and animals in preattentive vision. *Attention, Perception, & Psychophysics*, *63*(4), 676–697.

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*(1), 95–109.

Long, B., Störmer, V. S., & Alvarez, G. A. (2017). Mid-level perceptual features contain early cues to animacy. *Journal of Vision*, *17*(6):20, 1–20, https://doi.org/10.1167/17.6.20. [PubMed] [Article]

Perrett, D. I., Hietanen, J. K., Oram, M. W., Benson, P. J., & Rolls, E. T. (1992). Organization and functions of cells responsive to faces in the temporal cortex. *Philosophical Transactions of the Royal Society of London B*, *355*(1273), 23–30.

Perrinet, L. U., & Bednar, J. A. (2015). Edge co-occurrences can account for rapid categorization of natural versus animal images. *Scientific Reports*, *5*, 11400.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, *40*(1), 49–70.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439.

Schmidt, F., Hegele, M., & Fleming, R. W. (2017). Perceiving animacy from shape. *Journal of Vision*, *17*(11):10, 1–15, https://doi.org/10.1167/17.11.10. [PubMed] [Article]

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture account for rapid categorization. *Proceedings of the National Academy of Sciences, USA*, *104*(15), 6424–6429.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, *19*, 109–139.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55.

Thorpe, S., Fize, D., & Marlot, C. (1996, June 6). Speed of processing in the human visual system. *Nature*, *381*(6582), 520–522.

Tsao, D. Y., Freiwald, W. A., Tootell, R. B., & Livingston, M. S. (2006, February 3). A cortical region consisting entirely of face-selective cells. *Science*, *311*(5761), 670–674.

Yue, X., Pourladian, I. S., Tootell, R. B., & Ungerleider, L. G. (2014). Curvature-processing network in macaque visual cortex. *Proceedings of the National Academy of Sciences, USA*, *111*(33), E3467–E3475.