Attention, Perception, & Psychophysics (2025) 87:415–430
https://doi.org/10.3758/s13414-024-02976-6

Ensemble representation of animacy could be based on mid-level visual features

Natalia A. Tiurina[1] · **Yuri A. Markov**[2]

Accepted: 10 October 2024 / Published online: 4 December 2024

Abstract

Studies suggest that mid-level features could underlie object animacy perception. In the current research, we tested whether ensemble animacy perception is based on high- or mid-level features. We used five types of images of animals and inanimate objects: color, grayscale, silhouettes, texforms – unrecognizable images that preserve mid-level texture and shape information – and scrambled images. In the series of Experiments 1, we asked participants to evaluate the animacy of single images and sets of eight images using a 10-point scale. In the series of Experiments 2, participants were shown two sets of eight images and had to choose a more animate one in the two-alternative forced-choice (2AFC) task. We found that in both paradigms, observers could report the mean animacy of the set of texform images without direct access to information about high-level features. Thus, ensemble animacy could be extracted only based on mid-level features such as shape and texture without access to more high-level information.

Keywords Ensemble summary statistics · Animacy · Mid-level features

Introduction

The visual environment consists of numerous basic feature dimensions, which are naturally combined into various textures and patterns that can be recognized as meaningful objects. According to previous studies, all the properties of object representation could be divided into three levels: low-level, mid-level, and high-level features (Anderson, 2020; Freeman & Simoncelli, 2011; Groen, Silson, & Baker, 2017; Hayes & Henderson, 2021; Long et al., 2018; Peirce, 2015; Whitney & Yamanashi Leib, 2018). This is not a strict categorization of the features, but this hierarchical structure is linked to the processing of information from the lower (e.g., V1 connected with orientation processing; Hubel & Wiesel, 1959) to the higher (e.g., FFA, correlated with face processing; Kanwisher et al., 1997) visual areas. Basic visual features, like color or orientation, are usually considered low-level visual features, whereas high-level features are related to categorical or semantic representations. Mid-level features, which include information about texture and shape, are placed between them. Different levels of feature processing set different requirements for the visual system, which it cannot always meet due to its limitations (Cowan, 2001; Luck & Vogel, 1997; Mack & Rock, 1998; Miller, 1956; Pylyshyn & Storm, 1988; Simons & Chabris, 1999). However, we still perceive the world as rich, continuous, and detailed (Alvarez, 2011; Cohen et al., 2016). The visual system extracts statistical regularities from the information flow and represents them as summary statistics, which helps us in fast statistical estimation, rapid segmentation, and categorization of multiple objects (Alvarez, 2011; Ariely, 2001; Whitney & Yamanashi Leib, 2018).

Ensemble summary statistics aid us in making rapid statistical judgments about a set of objects without having full information about each of them (Ariely, 2001; Chong & Treisman, 2005; Whiting & Oriet, 2011). Early works in this area were focused primarily on the representation of low-level features such as visual motion (Watamaniuk & McKee, 1998; Watamaniuk et al., 1989), brightness (Bauer, 2009), orientation (Attarha & Moore, 2015; Dakin & Watt, 1997; Parkes et al., 2001), hue (Gardelle

Natalia A. Tiurina and Yuri A. Markov are contributed equally.

✉ Natalia A. Tiurina
nataliatiurina@gmail.com

[1] Department of Psychology, TUD Dresden University of Technology, Dresden, Germany

[2] Department of Psychology, Goethe University Frankfurt, Frankfurt am Main, Germany

& Summerfield, 2011; Maule & Franklin, 2015), and spatial location (Alvarez & Oliva, 2008). Recent studies have shown that observers can easily extract summary statistics for mid-level features such as object size (Ariely, 2001; Chong & Treisman, 2003; Haberman & Suresh, 2021; Markov & Tiurina, 2021; Tiurina & Utochkin, 2019), shape (Khayat et al., 2021; Sweeny et al., 2021), and texture (Cant & Xu, 2012, 2017; Koenderink et al., 2004). Observers could also estimate summary statistics for high-level visual features: emotion, gender, gaze direction, and head rotation (Florey et al., 2016, 2017; Haberman & Whitney, 2007, 2009; Han et al., 2020; Sweeny & Whitney, 2014; Yamanashi Leib et al., 2014); categorical characteristics (Khayat & Hochstein, 2019); economic value (Yamanashi Leib et al., 2020), and animacy (Yamanashi Leib et al., 2016).

However, it is still unclear whether high-level features are required to evaluate statistics for such complex abstract visual features as animacy or lifelikeness. Yamanashi Leib and colleagues (2016) demonstrated that observers were able to extract the mean animacy from a set of objects. Numerous experiments have provided strong evidence that this estimation occurred in spatial and temporal domains and could not be explained by subsampling or memory strategies. This finding suggested that this is an example of how a visual system could extract high-level and abstract information from multiple elements and that these results could not be explained by image properties or visual features of the stimuli (Whitney & Yamanashi Leib, 2018; Yamanashi Leib et al., 2016).

In contrast, studies suggest that estimation of *object* animacy can be based not on high-level features but rather on mid-level features, such as: second-order statistics (for different categories even beyond animacy: Torralba & Oliva, 2003), curvilinearity (Levin et al., 2001), mid-level texture and shape information (Li & Bonner, 2020; Long, Störmer, & Alvarez, 2017; Long et al., 2018; Schmidt et al., 2017; Wang et al., 2022; Zachariou et al., 2018), elongation and graspability (Almeida et al., 2014), and edge co-occurrences (Perrinet & Bednar, 2015). Importantly, these mid-level features could be accessed rapidly (Wang et al., 2022) and in a bottom-up visual processing way (Zachariou et al., 2018).

Access to each individual object representation is hardly possible during mean extraction from the set of items. How could a visual system extract and compress information about high-level properties without categorization of each individual item? In light of the findings in *object* animacy perception (e.g., Long et al., 2017, 2018), we investigated the role of high-level and mid-level features in *ensemble* animacy perception. The main goal of the study was not to find the exact features behind the ensemble animacy perception but to demonstrate that ensemble animacy perception is similar to object animacy perception and requires access only to mid-level features. In a series of experiments, we used five different types of stimuli: original color images, grayscale images with controlled luminance and contrast (SHINE toolbox; Willenbockel, et al., 2010), black silhouettes, texforms (Deza, Chen, Long, & Konkle, 2019; Freeman & Simoncelli, 2011; Long et al., 2018), and scrambled images created with diffeomorphic scrambling (Stojanoski & Cusack, 2014). The key manipulation was to test mean animacy extraction for texform images that contain some coarse texture and form information but are unrecognizable at the basic level – texforms lack high-level visual features, which are important for basic-level categorization (Deza, et al., 2019; Freeman & Simoncelli, 2011; Long, et al., 2017, 2018). Through all the experiments, we showed that observers were able to extract the mean animacy of the given set only based on mid-level features.

Experiment 1A

The series of Experiments 1 is based on the paradigms of previous studies investigating ensemble representation for complex visual properties (Han et al., 2020; Yamanashi Leib et al., 2016, 2020). In Experiment 1A, we tested all types of images. In Experiment 1B, we tested only texforms to avoid the influence of one type of image on another and recognition of texforms. In Experiment 1C, we tested whether participants used sampling strategies to perform the task. In Experiment 1A, observers completed two tasks: the object rating task (Fig. 1B) and the ensemble rating task (Fig. 1C). We presented one image (object rating task) or a set of images of animals and non-animals (ensemble rating task) and asked observers to report the animacy using the Likert scale. We used four types of images: original color images, grayscale images, black silhouettes, and texforms (Fig. 1A). Texforms contain the mid-level features of the original images; however, they are unrecognizable at the basic level. Thus, we used them as a "litmus test" to investigate whether ensemble animacy perception requires access to high-level information.

Method

Participants

Ten participants (three female, mean age = 25.3 years, SD = 6.3) were recruited through the Prolific platform (www.prolific.ac; Palan & Schitter, 2018; Peer et al., 2017) and given access to the online experiments using Pavlovia (https://pavlovia.org). The compensation for participation amounted to £5 per hour (Experiment 1A lasted approximately 25 min). The sample size was based on previous

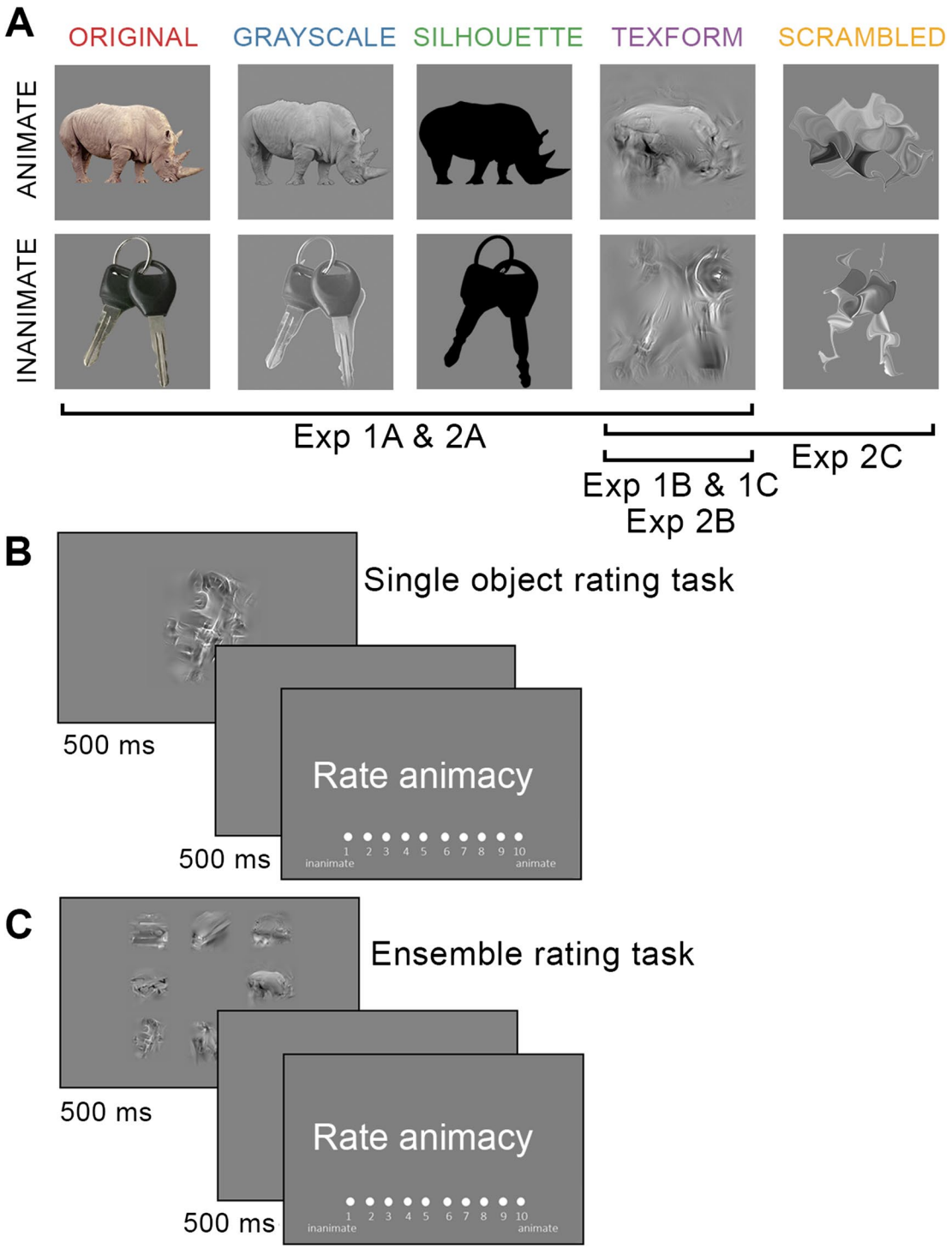

Fig. 1 Five types of stimuli were used in the Experiments (**A**). The time course of a typical trial in Experiment 1A, 1B, and 1C in the object rating task (**B**) and the ensemble rating task (**C**). In both tasks, participants had to rate animacy using the Likert scale. The sizes depicted here do not correspond to the real sizes of images in the experiment and serve illustrative purposes only

studies with a similar paradigm (Haberman & Whitney, 2011; Han et al., 2020; Sweeny et al., 2015; Yamanashi Leib et al., 2016). We conducted an additional power analysis using G*Power (Faul et al., 2007) that revealed a required sample size of six participants to detect an effect of 1.5 (based on the pilot data of texform object rating task, see

below) with a power of 0.80 and a two-sided t-test against a constant at an alpha level of 0.05. We recruited ten participants to avoid any technical problems with online experiments. Data analysis was also made on the subject level and we replicated our results in the following experiments and different paradigms (Experiment 1 vs. Experiment 2), thus sample size of ten participants is sufficient.

Apparatus and stimuli

The experiment was developed and presented online via PsychoPy v2020.2.10 (Bridges et al., 2020; Peirce et al., 2019).

In Experiment 1A, we used four types of stimuli images of animals and inanimate objects (Fig. 1A): original color images (Konkle & Caramazza, 2013), their high-contrast grayscale and texform versions, taken from subset of highly unrecognizable (on average, identified at the basic level less than 3% of the time) at the basic level texforms (Long et al., 2018), and black silhouettes of the same images we created for this study. All four types of stimuli were preliminarily tested in a separate online experiment ($N = 10$, two female, mean age = 23.3 years, $SD = 4.34$) with the object rating task in which participants rated images as animate and inanimate using the Likert scale from 1 to 10. The results of the test showed that images of animals ($M = 8.4$) were rated as significantly more animate than images of objects ($M = 1.7$; comparison: $t(9) = 28.892$, $p < 0.001$, Cohen's d = 9.136). Animacy ratings were different between the images of animals and inanimate objects for all types of images (original: animate ($M = 9.2$) versus inanimate ($M = 1.1$) – $t(9) = 175.103$, $p < 0.001$, Cohen's d = 55.372; grayscale: animate ($M = 9.45$) versus inanimate ($M = 1.16$) – $t(9) = 29.394$, $p < 0.001$, Cohen's d = 9.265; black silhouette: animate ($M = 8.89$) versus inanimate ($M = 1.26$) – $t(9) = 20.216$, $p < 0.001$, Cohen's d = 6.393; texform: animate ($M = 5.36$) versus inanimate ($M = 3.48$) – $t(9) = 4.801$, $p < 0.001$, Cohen's d = 1.518). The later effect size was taken for power analysis, since we expected a similar, if not slightly lesser, effect size could be anticipated for ensemble perception tasks. Considering the results of the online experiment, we selected the top 40% of the images of animals highly rated as animate and the top 40% of the images of objects highly rated as inanimate among all types of images to decrease the noise in the following experiments. The final stimuli set consists of 24 images per category and is available for open access at OSF (https://osf.io/7s5e2/).

Stimuli were presented on a gray background. The size of each image was 150 pix (approximately 3.8 degrees if the viewing distance was 60 cm and PPI was 96). In the *object rating task* (in Experiments 1A, B, and C and 2A, B, and C), one object was presented at the center of the screen. In the *ensemble rating task* (Experiments 1A, B, and C), the spatial arrangement of eight different images was snapped to the 3 × 3 grid, and the size of each cell was 200 × 200 pix (~ 5.05° × 5.05°). The central cell was always empty. Each image was assigned randomly to the center of the cells, and then a random number from -20 pix to 20 pix (~ 0.5°) was added to each of the coordinates.

Procedure

Each trial in the *object rating task* (Fig. 1B) started with a blank screen (500 ms), then one image appeared for 500 ms at the center of the screen, and after another 500 ms of blank screen, the response tool appeared. We asked participants to rate images as animate and inanimate using a scale from 1 to 10 where 1 corresponded to inanimate and 10 corresponded to animate. Participants always completed the object rating task before the ensemble rating task or two-alternative forced-choice (2AFC) task. In the object rating task, each image was presented once (96 trials for this task in total).

The *ensemble rating task* (Fig. 1C) was almost identical with only one difference – we showed a set of eight objects. We used nine possible configurations of the image set – from zero to eight animate images in the set. In Experiment 1A, each configuration for each type of stimuli (original, grayscale, black silhouettes, and texforms) was presented five times (45 trials per type of image, 180 trials in total).

Data analysis

In the *object rating task*, the mean rate was estimated separately for each type of image. The standard frequentist and Bayesian t-tests were calculated. The Bayesian t-test is a direct way to estimate evidence for H_1 against H_0 (Rouder et al., 2009). The Bayes factor (BF_{10}) was calculated using JASP 0.14.0.0 (JASP Team, 2021; Wagenmakers et al., 2017) and interpreted using the standard Jeffrey's scale (1961). The Cauchy distribution with a width of 0.707 was used as a prior distribution of effect sizes under H_0.

In the *ensemble rating task*, we estimated Pearson correlations for each type of image and each participant individually. We computed the correlations between participants' ratings in the ensemble rating task and the number of presented animate images in the set. Additionally, the "*mean individual rating*" was computed for each trial in the ensemble task by averaging the individual animacy ratings of each

object in the set. Individual animacy ratings were taken from the results of the object rating task for each participant separately. We also computed the correlation between "*mean individual rating*" and observers' answers in the ensemble rating task.

The statistics were calculated using JASP 0.14.0.0 (JASP Team, 2021) and Pingouin 0.3.11 Python Package (Vallat, 2018).

Results

Object rating task

The analysis of mean ratings in the *object rating task* data showed that images of animals were rated as more animate than images of inanimate objects regardless of the type of stimulus (Fig. 2A; image type $F(3, 27) = 11.205$, $p < 0.001$, $\eta^2_p = 0.555$; animacy $F(1, 9) = 1155.05$, $p < 0.001$, $\eta^2_p = 0.992$; interaction $F(3, 27) = 42.375$, $p < 0.001$, $\eta^2_p = 0.940$; original: animate ($M = 9.7$) versus inanimate ($M = 1.2$) – $t(9) = 40.144$, $p < 0.001$, $BF_{10} > 10^8$, Cohen's d = 12.695; grayscale: animate ($M = 9.59$) versus inanimate ($M = 1.22$) – $t(9) = 25.229$, $p < 0.001$, $BF_{10} > 10^6$, Cohen's d = 7.978; black silhouette: animate ($M = 9.37$) versus inanimate ($M = 1.21$) – $t(9) = 23.739$, $p < 0.001$, $BF_{10} > 10^6$, Cohen's d = 7.507; texform: animate ($M = 5.84$) versus inanimate ($M = 3.29$) – $t(9) = 9.397$, $p < 0.001$, $BF_{10} = 3229$, Cohen's d = 2.972).

Ensemble rating task

The analysis of the ensemble rating task was conducted in accordance with the analysis pipeline reported in previous studies (Han et al., 2020; Yamanashi Leib et al., 2016, 2020). We estimated two types of correlations: first, between observer ratings in the ensemble rating task and the *number of animate images*, and second, between observer ratings in the ensemble rating task and *mean individual rating* (see *General discussion*). As the pattern is consistent for both types of correlations, we describe them together in this and the following Experiments. We found strong correlations for original (*number of animate images*: Fisher's z = 1.45; Pearson's r = 0.88, $p < 0.001$; *mean individual rating*: Fisher's z = 1.44; Pearson's r = 0.87, $p < 0.001$), grayscale (*number of animate images*: Fisher's z = 1.39; Pearson's r = 0.87, $p < 0.001$; *mean individual rating*: Fisher's z = 1.37; Pearson's r = 0.87, $p < 0.001$), and black silhouette images (*number of animate images*: Fisher's z = 1.32; Pearson's r = 0.83, $p < 0.001$; *mean individual rating*: Fisher's z = 1.32; Pearson's r = 0.83, $p < 0.001$). The correlations for texforms were lower but still reliably high across all participants (*number of animate images*: Fisher's z = 0.61; Pearson's r = 0.53, $p < 0.01$; *mean individual rating*: Fisher's z = 0.54; Pearson's r = 0.48, $p < 0.01$). These correlations suggest that observers were able to report mean animacy for all types of images, with decreased performance for texform images. The averaged correlations across observers are presented in Fig. 3A, with individual correlations available in supplementary materials (Online Supplementary Materials (OSM), Figs. 1 and 2).

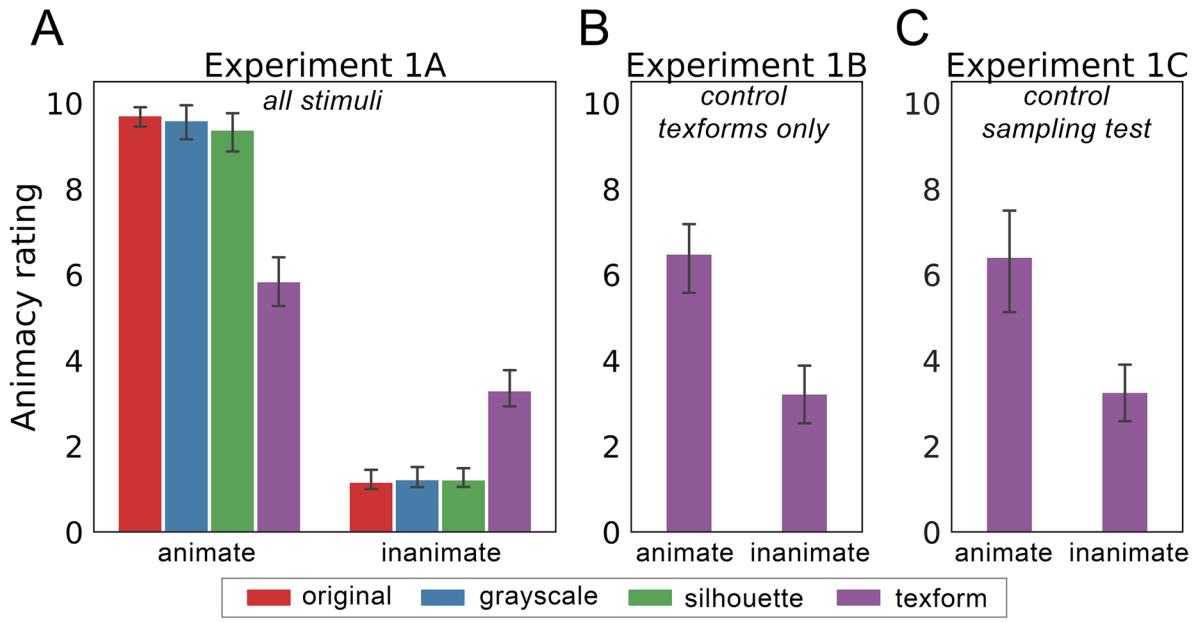

Fig. 2 The results of the object rating task for Experiment 1A, Experiment 1B, and Experiment 1C. The higher animacy rating value indicates that observers rated the image as more animate. Error bars depict 95% confidence intervals

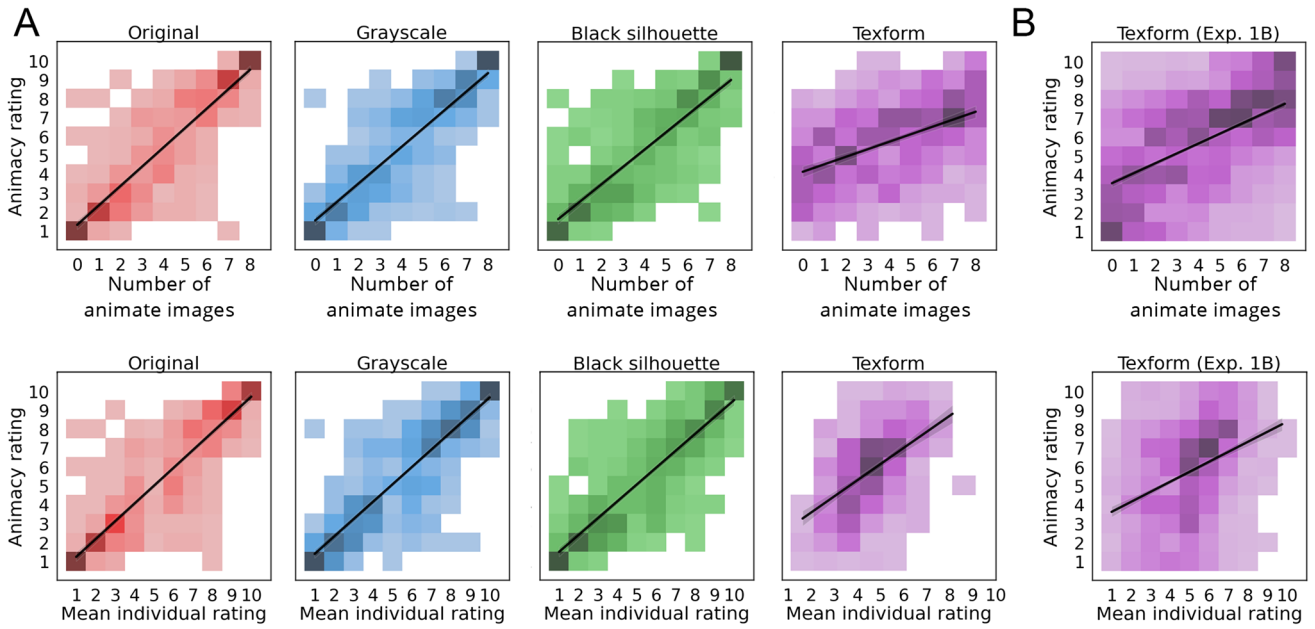

Fig. 3 The results for the ensemble rating task for Experiment 1A and Experiment 1B. Correlations across all observers were estimated for four types of images. Darker colors indicate a higher number of observations that fall within discrete bins. The black line illustrates the regression model fit

Experiment 1B

Since we used a small subset of stimuli and each image was presented in all four types, we admit the possibility that texforms could be recognized in the object rating task through other types of the same images. Thus, we conducted Experiment 1B in which only texforms were used as stimuli in both object and ensemble rating tasks.

Method

Participants

Ten participants (three female, mean age = 24.8 years, SD = 4.81) were recruited through the Prolific platform (www.prolific.ac; Palan & Schitter, 2018; Peer et al., 2017) and given access to the online experiments using Pavlovia (https://pavlovia.org). The compensation for participation amounted to £5 per hour (Experiment 1B – 12 min).

Apparatus, stimuli, procedure, and data analysis

A similar apparatus, stimuli, procedure, and data analysis were used in Experiment 1B as in Experiment 1A. The main difference from Experiment 1A was that only texform images were used in both the *object rating task* and *ensemble rating task*. In Experiment 1B, the number of trials for the ensemble task amounted to 180 (20 trials per configuration).

Results

Object rating task

We found that the texform images of animals ($M = 6.5$) were rated as more animate than the texform images of inanimate objects (Fig. 2B; $M = 3.2$; comparison: $t(9) = 8.831$, $p < 0.001$, $BF_{10} = 2075$, Cohen's d = 2.792).

Ensemble rating task

We established significant correlations between observer ratings in the ensemble rating task and *the number of animate images* (Fig. 3B, Fisher's z = 0.65; Pearson's r = 0.55, $p < 0.001$) as well as between the observer ratings and *mean individual rating* (Fisher's z = 0.57; Pearson's r = 0.50, $p < 0.001$). This suggests that observers were able to extract mean animacy even without cues from other types of objects.

Experiment 1C

In Experiment 1C, we sought to verify that the results of Experiments 1A and 1B are based on the processing of a set of images rather than one random object from the set. The results of a previous study (Yamanashi Leib et al., 2016) showed that observers were able not only to report the animacy of one random object but also to integrate information

about animacy from the set of color images. We had to confirm similar effects for texforms. We always manipulated the animacy of the set by changing the proportion of animate and inanimate items; thus, the probability of randomly choosing one animate object from the set correlated with the number of animate images on display.

We estimated correlations between observer ratings and the *number of animate images* as well as between observer ratings and *mean individual rating* in the same way as we computed them in Experiments 1A and 1B. After that, we compared the strength of these correlations between the conditions where all eight images were presented (the *whole set condition*) and those where only a subset of images was shown (*subset conditions*). If correlations in subset conditions did not differ from the whole set condition, it would indicate the absence of a beneficial effect of the additional information relevant to the task performance (Fig. 4A). Hence, we could conclude that observers used the subsampling strategy. However, if the strengths of these correlations differed across conditions with the highest correlation in the whole set condition, we could assume that observers extract animacy by taking into account at least more than one item (Fig. 4B).

We presented a subset – one, two, four, or six objects randomly chosen from the set – or the whole set of eight objects and asked observers to report the mean animacy of the display. For instance, imagine the initial whole set consisted of three inanimate and five animate objects. In the subset condition, we randomly hid six items, and only two objects – one animate and one inanimate – stood visible for observers to report.

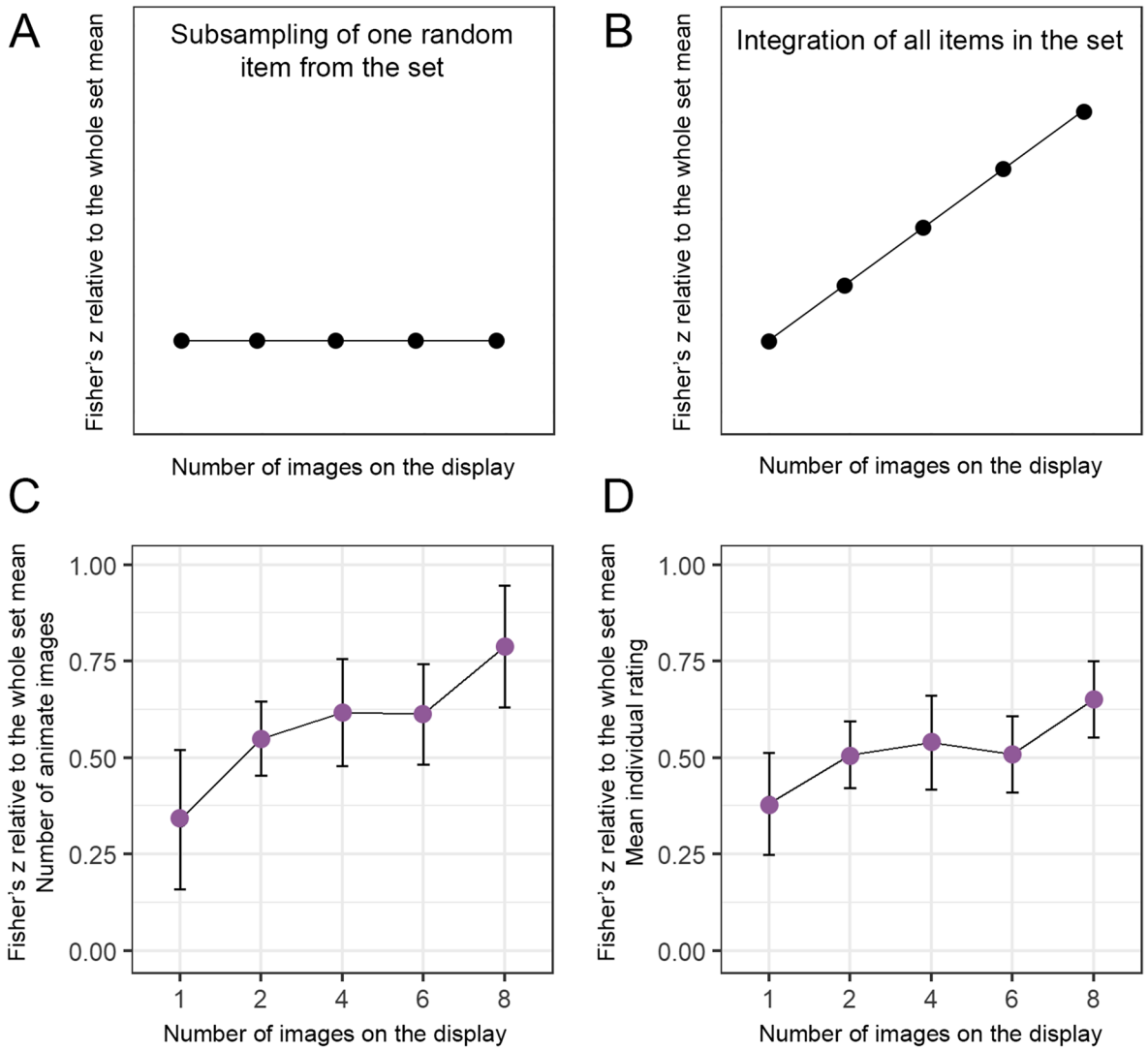

Fig. 4 (**A**) Prediction for one random object sampling from a set: Fisher's z values would not increase, even as more information became available. (**B**) Prediction for integration of all items in a set: Fisher's z values would systematically increase with the number of images presented. (**C**) Fisher's z for the number of animate images and (**D**) Fisher's z for mean individual rating estimated for the number of images presented

Then, we correlated the reported mean animacy with the mean animacy of the whole set (five animate images). As in previous experiments, we estimated the correlation between three measures: observer ratings, the *number of animate images,* and *mean individual rating*. Importantly, the *number of animate images* and *mean individual rating* were calculated as if *the whole set was visible*. This method has already been described in previous studies (Han et al., 2020; Piazza, et al., 2013; Sweeny et al., 2013, 2014, 2015; Wolfe et al., 2015; Yamanashi Leib et al., 2014, 2016, 2020), allowed us to simulate various sampling strategies and estimate their performance.

Method

Participants

Ten participants (four female, mean age = 22.2 years, SD = 3.91) were recruited through the Prolific platform (www.prolific.ac; Palan & Schitter, 2018; Peer et al., 2017) and given access to the online experiments using Pavlovia (https://pavlovia.org). The compensation for participation amounted to £5 per hour (Experiment 1C – 18 min).

Apparatus, stimuli, procedure, and data analysis

A similar apparatus, stimuli, procedure, and data analysis were used in Experiment 1C as in Experiments 1A and 1B. The main differences were that we used only texform images and also presented different subset conditions. We presented the whole set or only six, four, two, or one random image from the set to the observers. The number of trials for the ensemble task amounted to 225 (45 trials per "subset"). The whole set condition was identical to the texform condition in Experiment 1A. Also, in Experiment 1C, the *number of animate images* and *mean individual rating* for subset conditions were calculated as if *the whole set was visible.*

Results

Object rating task

We found that the texform images of animals ($M = 6.408$) were rated as more animate than the texform images of inanimate objects (Fig. 2C; $M = 3.3$; comparison: $t(9) = 6.092$, $p < 0.001$, $BF_{10} = 172.3$, Cohen's d = 1.927). The ratings of texforms were similar to the data obtained in Experiment 1A and Experiment 1B.

Ensemble rating task

As in Experiments 1A and 1B, we observed significant correlations for texforms in the whole set condition (*number of animate images*: Fisher's z = 0.79; Pearson's r = 0.62, $p < 0.005$; *mean individual rating*: Fisher's z = 0.65; Pearson's r = 0.56, $p < 0.001$). Next, we compared correlations (Fisher's z) for each subset condition using RM ANOVA (Fig. 4C and D). We established a linear trend: Fisher's z values increased with the number of images on display (*number of animate images*: $F(4, 36) = 6.521$, $p < 0.001$, $BF_{10} = 71.3$, $\eta^2 = 0.420$; *mean individual rating*: $F(4, 36) = 4.035$, $p = 0.008$, $BF_{10} = 6.42$, $\eta^2 = 0.310$), indicating that observers used additional relevant information and did not base their answers on one randomly chosen item.

Discussion

In Experiments 1A, 1B, and 1C, we found strong correlations between observers' reports and the number of images of animate objects on the screen. The results for original images were similar to the results reported in a previous study by Yamanashi Leib and colleagues (2016). Similarly, high correlations for black silhouettes and grayscale images suggested that mean animacy estimation was not hindered by the absence of particular features – textures and colors. Indeed, we observed moderate correlations for texform images. According to previous studies, correlations in the ensemble rating tasks depend on the noise level of the image (Han et al., 2020) and the inconsistency of observers' answers (Yamanashi Leib et al., 2020). Texform images are noisy; therefore, it is significantly harder to rate them even in the object rating task. Thus, it is incorrect to compare texform images and other types of images in the performance. However, significant and moderate correlations for texform images allow us to conclude that observers were able to report the mean animacy of a set. In Experiment 1B, observers saw only texforms images and not the original ones; thus, we ruled out any possible identification of texforms but still found significant correlations. In Experiment 1C, we tested whether observers used a subsampling strategy. In line with previous studies (e.g., Yamanashi Leib et al., 2016), correlations were lower in subset conditions than in the whole set condition. Fisher's z values systematically increased with the number of images presented, suggesting that observers were able to integrate information from all the images on the screen. Thus, our results cannot be explained by subsampling strategies.

Experiment 2A

In the series of Experiments 2, we used the 2AFC paradigm (Whitney & Yamanashi Leib, 2018) to confirm that our results are not specific to one paradigm. 2AFC paradigm eliminates biases inherent in Likert scales and various observers' heuristics that could occur with the rating scales (Böckenholt, 2017; Huang, 2016). Similarly, to the series of Experiment 1, each observer first reported the animacy

of individual objects in the object rating task and after performed the ensemble 2AFC task. In Experiment 2A, we tested all types of images. In Experiment 2B, we tested only texforms. In Experiment 2C, we compared texforms with scrambled images. We asked participants to compare two sets of images and determine which set had higher mean animacy. We manipulated the difference in the number of animate images between sets (from one to six images). In other aspects, the logic of Experiment 2A was similar to that of Experiment 1A.

Method

Participants

Ten participants (three female, mean age = 26.1 years, SD = 6.03) were recruited through the Prolific platform (www.prolific.ac; Palan & Schitter, 2018; Peer et al., 2017) and given access to the online experiments using Pavlovia (https://pavlovia.org). The compensation for participation amounted to £5 per hour (Experiment 2A – 45 min).

Apparatus and stimuli

The experiment was developed and presented online via PsychoPy v2020.2.10 (Bridges et al., 2020; Peirce et al., 2019). In Experiment 2A, we used all four types of stimuli images of animals and inanimate objects: original, grayscale, black silhouette, texforms (Fig. 1A). The *object rating task* was the same as in Experiment 1A and observers participated in it before the ensemble *2AFC task*.

Stimuli were presented on a gray background. The size of each image was 150 pix (approximately 3.8 degrees if the viewing distance was 60 cm and PPI was 96). In the *2AFC task*, 16 different images were presented in two spatially separated sets (sizes of the images and spatial arrangement of each set were the same as in the *ensemble rating task*). One set was located 400 pix (~ 10.08°) to the left from the center of the screen, and the second set was located 400 pix (~ 10.08°) to the right.

Procedure

In the *2AFC task*, the difference in the number of animate images between the two sets varied from one to six images. Two sets were presented for 500 ms, followed by 500 ms of blank screen. After that, observers pressed the left or right button to report which set had higher mean animacy (Fig. 5). Feedback was displayed for 1 s after each trial. In Experiment 2A, there were 720 trials in total (30 trials per difference and type of image).

Data analysis

In the *object rating task*, the mean rate was estimated separately for each type of image. The standard frequentist and Bayesian t-tests were calculated.

In the *2AFC task*, we estimated the percentage of correct answers for each condition and each difference in the number of animate images between the two sets. The standard frequentist and Bayesian t-tests and RM ANOVAs were

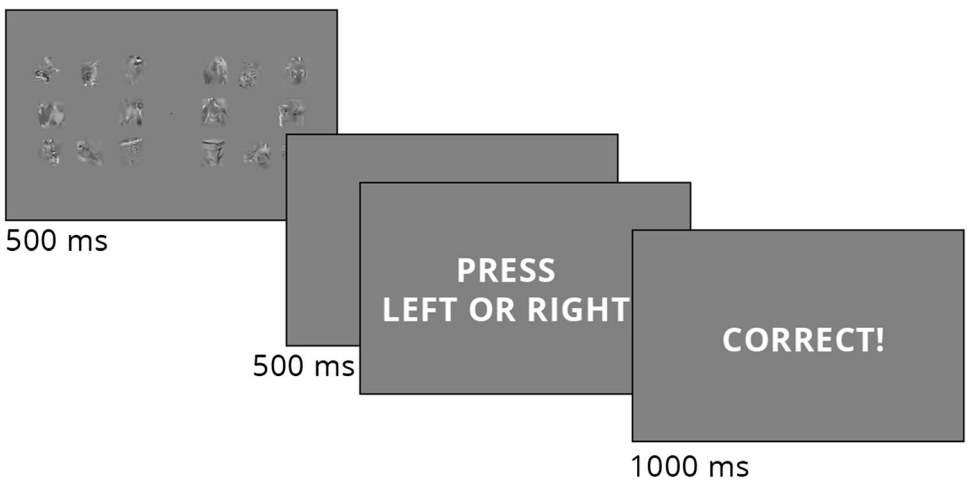

Fig. 5 The time course of a typical trial in Experiments 2A, 2B, and 2C in the two-alternative forced-choice (2AFC) task (**A**). Participants had to compare two sets and choose a more animate one by pressing the left and right buttons. The sizes depicted here do not correspond to the real sizes of images in the experiment and serve illustrative purposes only

performed. A Bonferroni correction was made for multiple comparisons in calculating the statistical significance level.

The statistics were calculated using JASP 0.14.0.0 (JASP Team, 2021) and Pingouin 0.3.11 Python Package (Vallat, 2018).

Results

Object rating task

All types of images of animals were rated as more animate than the images of inanimate objects (Fig. 6A, image type $F(3, 27) = 4.852$, $p = 0.008$, $\eta^2_p = 0.350$; animacy $F(1, 9) = 2357.46$, $p < 0.001$, $\eta^2_p = 0.996$; interaction $F(3, 27) = 143.6$, $p < 0.001$, $\eta^2_p = 0.941$; – original: animate ($M = 9.94$) versus inanimate ($M = 1.06$) – $t(9) = 149.27$, $p < 0.001$, $BF_{10} > 10^{12}$, Cohen's d = 47.203; grayscale: animate ($M = 9.98$) versus inanimate ($M = 1.06$) – $t(9) = 189.486$, $p < 0.001$, $BF_{10} > 10^{13}$, Cohen's d = 59.921; black silhouette: animate ($M = 9.65$) versus inanimate ($M = 1.242$) – $t(9) = 48.728$, $p < 0.001$, $BF_{10} > 10^{9}$, Cohen's d = 15.409; texform: animate ($M = 6.79$) versus inanimate ($M = 3.329$) – $t(9) = 7.647$, $p < 0.001$, $BF_{10} = 764$, Cohen's d = 2.418).

2AFC task

We estimated the percentage of correct answers for each type of image and each difference in the number of animate images. We found a significant effect of the type of image (Fig. 7A; $F(3,27) = 47.613$, $p < 0.001$, $BF_{10} > 10^5$, $\eta^2_p = 0.841$). The percentage of correct answers was lower for texform images ($M = 0.757$) in comparison with all other types of images (original $M = 0.882$; comparison: $t(9) = 11.080$, $p_{holm} < 0.001$, $BF_{10} > 10^{13}$, Cohen's d = 3.504; grayscale $M = 0.856$; comparison: $t(9) = 8.804$, $p_{holm} < 0.001$, $BF_{10} > 10^{8}$, Cohen's d = 2.784; silhouette $M = 0.854$; comparison: $t(9) = 8.606$, $p_{holm} < 0.001$, $BF_{10} > 10^{7}$, Cohen's d = 2.722). The effect of the difference in the number of animate images between the two sets was significant – the percentage of correct answers was higher in trials with larger differences between the two sets ($F(5,45) = 111.231$, $p < 0.001$, $BF_{10} > 10^{46}$, $\eta^2_p = 0.925$). We did not find any significant interaction between the two factors ($F(15,135) = 1.088$, $p = 0.373$, $BF_{10} = 0.08$, $\eta^2_p = 0.108$).

The percentage of correct answers for all types of images differed from the 50% guessing threshold (original: $t(9) \geq 7.446$, $p \leq 0.001$, $BF_{10} \geq 638$, Cohen's d ≥ 2.355; grayscale: $t(9) \geq 4.440$, $p \leq 0.002$, $BF_{10} \geq 27$, Cohen's d ≥ 1.404; silhouette: $t(9) \geq 5.674$, $p \leq 0.001$, $BF_{10} \geq 111$, Cohen's d ≥ 1.794; texform: $t(9) \geq 4.200$, $p \leq 0.002$, $BF_{10} \geq 20$, Cohen's d ≥ 1.328). This suggests that observers were able to perform the task for all types of images.

Experiment 2B

In this experiment, we tested whether recognizable types of images (original color, grayscale, and silhouettes) can provide cues for texform recognition in the 2AFC task. To eliminate any familiarity effects, we used only texforms

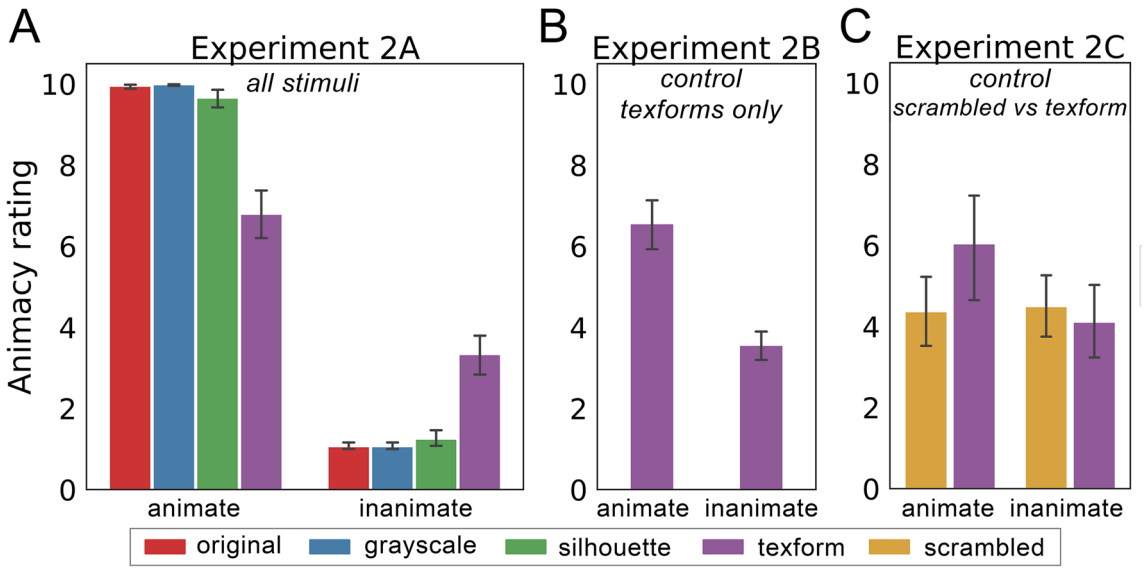

Fig. 6 The results for the object rating task for Experiment 2A, Experiment 2B, and Experiment 2C. A higher animacy rating value indicates that observers rated the image as more animate. Error bars depict 95% confidence intervals

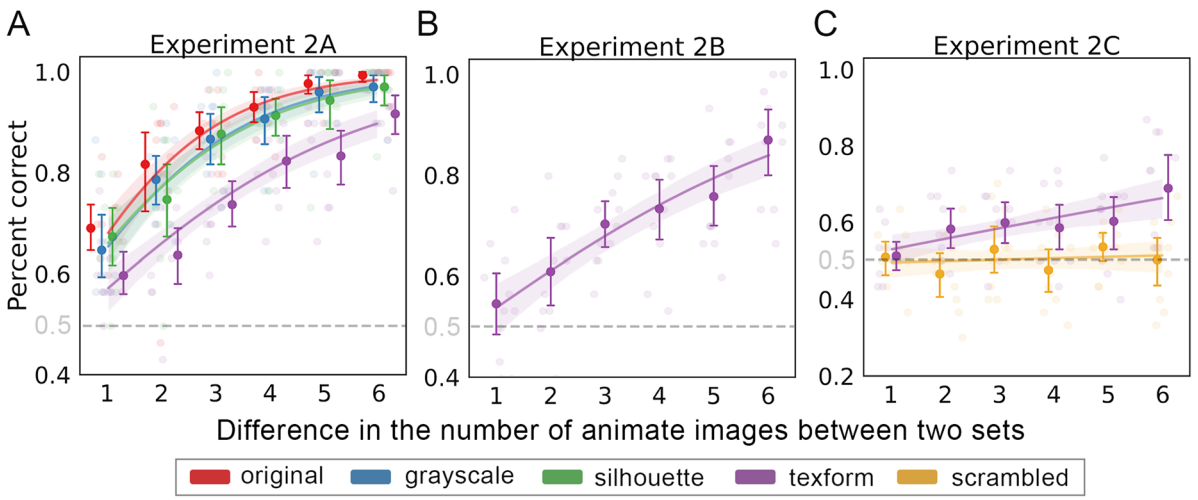

Fig. 7 The results for the two-alternative forced-choice (2AFC) task for Experiment 2A, Experiment 2B, and Experiment 2C. The lines of logistic regression fits are used for illustrative purposes. Transparent circles indicate individual observer data per condition. The dashed transparent line illustrates a 50% guess level. Error bars depict 95% confidence intervals

in this experiment. We asked participants to compare two sets of images and determine which set had higher mean animacy. We manipulated the difference in the number of animate images between sets (from one to six images). In other aspects, the logic of Experiment 2B was similar to that of Experiment 1B.

Method

Participants

Eleven participants (one female, mean age = 22.27 years, SD = 5.07) were recruited through the Prolific platform (www.prolific.ac; Palan & Schitter, 2018; Peer et al., 2017) and given access to the online experiments using Pavlovia (https://pavlovia.org). The compensation for participation amounted to £5 per hour (Experiment 2B – 20 min).

Apparatus, stimuli, procedure, and data analysis

A similar apparatus, stimuli, procedure, and data analysis were used in Experiment 2B as in Experiment 2A. The main difference from Experiment 2A was that only texform images were used in both the *object rating task* and the *2AFC task*. There were 180 trials in Experiment 2B in the *2AFC task*.

Results

Object rating task

We found that the texform images of animals ($M = 6.55$) were rated as more animate than the texform images of objects ($M = 3.56$; comparison: $t(10) = 8.512$, $p < 0.001$, $BF_{10} = 2970$, Cohen's d = 2.566; Fig. 6B).

2AFC task

We found the significant effect of the difference in the number of animate images between the two sets (Fig. 7B; $F(5,50) = 22.831$, $p < 0.001$, $BF_{10} > 10^9$, $\eta^2_p = 0.695$). The percentage of correct answers for texforms was higher than 50% in all conditions ($t(10) \geq 3.157$, $p \leq 0.01$, $BF_{10} \geq 5.955$, Cohen's d ≥ 0.952), except for the condition where the difference between two sets was one animate image: $t(10) = 1.386$, $p = 0.196$, $BF_{10} = 0.638$, Cohen's d = 0.418).

Experiment 2C

All previous experiments provided evidence that ensemble animacy could be estimated with mid-level features. The usage of texforms does not exclude the possibility that not only mid-level but also low-level features underlie ensemble animacy perception. In Experiment 2C, we compared texforms that retain both mid-level and low-level features with scrambled images which had information only about low-level features (Stojanoski & Cusack, 2014). Scrambled images are completely unrecognizable (which was supported by the results of the object rating task of this experiment) and contain highly distorted information about the shape of the original image. Thus, they are perfect candidates to test whether low-level features could underlie ensemble animacy perception.

Method

Participants

Ten participants (three female, mean age = 24.0 years, SD = 5.85) were recruited through the Prolific platform (www.prolific.ac; Palan & Schitter, 2018; Peer et al., 2017) and given access to the online experiments using Pavlovia (https://pavlovia.org). The compensation for participation amounted to £5 per hour (Experiment 2C – 30 min).

Apparatus, stimuli, procedure, and data analysis

A similar apparatus, stimuli, procedure, and data analysis were used in Experiment 2C as in Experiment 2B. The main difference from Experiment 2C was that a new type of image – scrambled images – was added and used in both the *object rating task* and the *2AFC task*. Scrambled images (Fig. 1A) were created from grayscale images using the diffeomorphic transformation package for Matlab with 20 steps and a maximum distortion of 80 (Stojanoski & Cusack, 2014; https://github.com/rhodricusack/diffeomorph). The final stimuli set consists of 24 images per category and is available for open access via the Open Science Framework (https://osf.io/7s5e2/). There were 360 trials (180 for each type of image) in Experiment 2C in the *2AFC task*.

Results

Object rating task

We didn't find a significant effect of the type of image ($F(1, 9) = 2.278$, $p = 0.165$, $\eta^2_p = 0.202$), but we found a significant effect of animacy ($F(1, 9) = 18.089$, $p = 0.002$, $\eta^2_p = 0.668$) and interaction between animacy and the type of the image ($F(1, 9) = 20.375$, $p = 0.001$, $\eta^2_p = 0.694$). The texform images of animals ($M = 6.03$) were rated as more animate than the texform images of objects ($M = 4.104$; comparison: $t(9) = 4.667$, $p < 0.001$, $BF_{10} = 35.76$, Cohen's d = 1.476). However, ratings for the scrambled images of animals ($M = 4.36$) and the scrambled images of objects ($M = 4.49$) did not differ from each other ($t(9) = 0.831$, $p = 0.428$, $BF_{10} = 0.411$, Cohen's d = 0.263; Fig. 6C).

2AFC task

We found a significant effect of the type of image. The percentage of correct answers was higher for texform images ($M = 0.596$) in comparison with scrambled images ($M = 0.504$; Fig. 7C; comparison: $F(1,9) = 19.189$, $p = 0.002$, $BF_{10} > 10^5$, $\eta^2_p = 0.681$). The effect of the difference in the number of animate images between the two sets was not significant ($F(5,45) = 2.221$, $p = 0.069$, $BF_{10} = 0.514$, $\eta^2_p = 0.198$). The interaction between two factors was significant ($F(15,135) = 3.062$, $p = 0.018$, $BF_{10} = 1.4$, $\eta^2_p = 0.254$).

The percentage of correct answers for scrambled images did not differ from the 50% guessing threshold ($t(9) \leq 1.766$, $p \geq 0.111$, $BF_{10} \leq 0.978$, Cohen's d ≤ 0.559). The percentage of correct answers for texform images was higher than 50% ($t(9) \geq 2.685$, $p \leq 0.025$, $BF_{10} \geq 3.002$, Cohen's d ≥ 0.849), except for the differences of one image between two sets – $t(9) = 0.647$, $p = 0.534$, $BF_{10} = 0.369$, Cohen's d = 0.205). Thus, we discovered that observers were unable to report mean animacy based only on low-level features.

Discussion

In Experiments 2A, 2B, and 2C, we showed, using the 2AFC paradigm, that observers could determine which of the two sets had higher animacy. In general, the percentage of correct answers for texforms images was lower compared to original color, grayscale, and silhouette images. The pattern observed for texforms images was the same as for other types of images: the percentage of correct answers systematically increased with the rise of the difference in the number of animate images between two sets. Importantly, observers were relatively precise in discriminating between two sets of texform images. We did not find any effect for scrambled images; thus, our results cannot be explained by simple differences in low-level features between animate and inanimate images.

General discussion

In line with previous studies (Yamanashi Leib et al., 2016), we demonstrated that observers could extract the mean animacy from a set. Crucially, we showed that mean animacy judgments could be made without complete access to high-level information; observers can distinguish between more and less animated sets of texforms, even though this type of image preserves information only about low- and mid-level features. Furthermore, in two different experimental paradigms, we showed that subsampling strategies fail to explain our results and demonstrated that low-level features themselves are unable to support precise animacy judgments. Ensemble animacy perception is based on mid-level features, similar to object animacy perception (e.g., Long et al., 2017, 2018). Taken together, these findings suggest that our visual system can compress complex abstract information (e.g., animacy) from several objects to overcome limitations (Cowan, 2001; Luck & Vogel, 1997; Mack & Rock, 1998; Miller, 1956; Pylyshyn & Storm, 1988; Simons & Chabris, 1999).

It is important to note that performance for texform images was lower in ensemble tasks than performance for other types of images (except scrambled images). However, the same decrease could be seen in object rating tasks as well. Being highly noisy, texform images are hard to compare with other types of images. Nevertheless, information left in the texform images was enough for observers to report animacy.

A general explanation of how we compute summary statistics frequently involves pooling models: information about visual features is pooled from the lower areas of the visual cortex, "mixed up", and averaged in the higher areas (Balas et al., 2009; Haberman & Whitney, 2011; Parkes et al., 2001; Robinson & Brady, 2023; Rosenholtz et al., 2012; Utochkin et al., 2024; Whitney & Yamanashi Leib, 2018). However, some studies (Whitney & Yamanashi Leib, 2018; Yamanashi Leib et al., 2016) suggested that pooling models are insufficient to explain such high-level ensemble properties as mean animacy, assuming a unique mechanism behind this process. However, in light of our findings, we argue that animacy ensemble perception could be based on mid-level features extraction and could use the same mechanism as other types of low- or mid-level features, such as orientations or textures. Visual features associated with animacy could be pooled from lower visual regions to the higher visual areas (e.g., occipitotemporal cortex, which is associated with animacy processing (Konkle & Caramazza, 2013; Long et al., 2018) where animacy judgments based on the population response could be made.

The visual features related to animacy are still under discussion. In our study, we rather arbitrarily use mid-level feature terminology, because the goal of the study was not to find the exact features underlying the ensemble animacy perception, but rather to demonstrate that basic recognition is not required for this. However, from our results and previous studies we could hypothesize what features are important in animacy judgments. Curvature is one of the mid-level features, mainly discussed in the literature as being associated with object animacy (Levin et al., 2001; Long et al., 2017, 2018; Schmidt et al., 2017; Zachariou et al., 2018): the shape of inanimate objects varies from boxy to curvy, while the shape of animals tends to be curvier. A recent study argues that observers could make animacy judgments about objects based only on curvature (Zachariou et al., 2018). Ensemble summary statistics studies demonstrated that observers could extract the mean shape, which could vary from very boxy to very curvy (Robinson & Brady, 2023), as well as mean texture (Koenderink et al., 2004) or other types of shape information (Khayat et al., 2021; Sweeny et al., 2021). Thus, curvature is a feature that may be strongly related to the object and ensemble animacy perception. Texforms images preserve curvature statistics, confirming the hypothesis that curvature is an important feature. Our results for black silhouettes also could confirm this hypothesis, because the performance for them was on the same level as for original or grayscale images. However, they could be easily recognized and it is impossible in this case to disentangle the contribution of curvature and familiarity. Additional studies are required to explore the origin features behind animacy perception of both single objects and ensembles.

Our results demonstrate that mid-level features alone can be used to predict the mean animacy of a set. We could speculate whether other complex features of the set – for example, face properties or economic value (Haberman & Whitney, 2007, 2009; Han et al., 2020; Yamanashi Leib et al., 2020) – could be explained by mid-level features. For example, research has shown that face processing could be based only on low-level visual features (Becker et al., 2011; Coelho et al., 2010; Purcell & Stewart, 2010). Moreover, observers can extract mean face properties from a set of inverted faces (Elias et al., 2017; Haberman & Whitney, 2009; Sweeny & Whitney, 2014), which have the same low-level features as upright faces but are processed less "holistically" (Farah et al., 1995; Farah et al., 1998; but see Richler et al., 2011). This could indicate that holistic or high-level processing is not required for mean estimation, suggesting that these judgments could be based on low or mid-level features. A recent study by Han and colleagues (2020) demonstrates that mean emotional valence can be estimated for Mooney faces which cannot be recognized by separate features. These results leave open the question of whether the processing of several faces could be based purely on the low- or mid-level features. We also do not rule out the possibility that high-level information could contribute to the statistical decisions about abstract dimensions; however, the exact contribution of different types of features should be examined in future studies.

The visual environment contains a lot of regularities and correlations, and our very limited visual system smartly detects and uses them, creating rich and continuous percepts (Alvarez, 2011; Cohen et al., 2016). To understand the nature of visual processing, all levels of the visual hierarchy must be investigated. Our results suggest that the visual system can extract summary statistics about the abstract dimension from a set, bypassing the high-level processing of individual objects.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13414-024-02976-6.

Acknowledgements The authors thank Timothy F. Brady and anonymous reviewers for their helpful comments.

Funding This study was supported by the Swiss Government Excellence Scholarship awarded by the Federal Commission for Scholarships for Foreign Students FCS (Natalia Tiurina).

Data availability The data and stimuli from all the experiments reported in this article can be accessed at: https://osf.io/7s5e2/

Code availability Not applicable.

Declarations

Ethics approval The study was conducted in accordance with the Declaration of Helsinki (World Medical Organization, 2013). This study was not preregistered.

Consent to participate Not applicable.

Consent for publication Not applicable.

Conflicts of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Almeida, J., Mahon, B. Z., Zapater-Raberov, V., Dziuba, A., Cabaço, T., Marques, J. F., & Caramazza, A. (2014). Grasping with the eyes: The role of elongation in visual recognition of manipulable objects. *Cognitive, Affective and Behavioral Neuroscience, 14*(1), 319–335. https://doi.org/10.3758/s13415-013-0208-0

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences, 15*(3), 122–131. https://doi.org/10.1016/j.tics.2011.01.003

Alvarez, G. A., & Oliva, A. (2008). The Representation of Simple Ensemble Visual Features Outside the Focus of Attention. *Psychological Science, 19*(4), 392–398. https://doi.org/10.1111/j.1467-9280.2008.02098.x

Anderson, B. L. (2020). Mid-level vision. *Current Biology, 30*(3), R105–R109. https://doi.org/10.1016/j.cub.2019.11.088

Ariely, D. (2001). Seeing Sets: Representation by Statistical Properties. *Psychological Science, 12*(2), 157–162. https://doi.org/10.1111/1467-9280.00327

Attarha, M., & Moore, C. M. (2015). The capacity limitations of orientation summary statistics. *Attention, Perception & Psychophysics, 77*(4), 1116–1131. https://doi.org/10.3758/s13414-015-0870-0

Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision, 9*(12), 13–13. https://doi.org/10.1167/9.12.13

Bauer, B. (2009). Does Stevens's power law for brightness extend to perceptual brightness averaging? *The Psychological Record, 59*(2), 171–185. https://doi.org/10.1007/BF03395657

Becker, D. V., Anderson, U. S., Mortensen, C. R., Neufeld, S. L., & Neel, R. (2011). The face in the crowd effect unconfounded: Happy faces, not angry faces, are more efficiently detected in single- and multiple-target visual search tasks. *Journal of Experimental Psychology: General, 140*(4), 637–659. https://doi.org/10.1037/a0024060

Böckenholt, U. (2017). Measuring response styles in likert items. *Psychological Methods, 22*(1), 69–83. https://doi.org/10.1037/met0000106

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ*, *8*. https://doi.org/10.7717/peerj.9414

Cant, J. S., & Xu, Y. (2012). Object Ensemble Processing in Human Anterior-Medial Ventral Visual Cortex. *Journal of Neuroscience, 32*(22), 7685–7700. https://doi.org/10.1523/JNEUROSCI.3325-11.2012

Cant, J. S., & Xu, Y. (2017). The contribution of object shape and surface properties to object ensemble representation in anterior-medial ventral visual cortex. *Journal of Cognitive Neuroscience, 29*(2), 398–412. https://doi.org/10.1162/jocn_a_01050

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research, 43*(4), 393–404. https://doi.org/10.1016/S0042-6989(02)00596-5

Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. *Vision Research, 45*(7), 891–900. https://doi.org/10.1016/j.visres.2004.10.004

Coelho, C. M., Cloete, S., & Wallis, G. (2010). The face-in-the-crowd effect: When angry faces are just cross (es). *Journal of Vision, 10*(1), 7. https://doi.org/10.1167/10.1.7

Cohen, M. A., Dennett, D. C., & Kanwisher, N. (2016). What is the Bandwidth of Perceptual Experience? *Trends in Cognitive Sciences, 20*(5), 324–335. https://doi.org/10.1016/j.tics.2016.03.006

Cowan, N. (2001). The magical number 4 in short term memory. A reconsideration of storage capacity. *Behavioral and Brain Sciences*, *24*(4), 87–186. https://doi.org/10.1017/S0140525X01003922

Dakin, S. C., & Watt, R. J. (1997). The computation of orientation statistics from visual texture. *Vision Research, 37*(22), 3181–3192. https://doi.org/10.1016/S0042-6989(97)00133-8

De Gardelle, V., & Summerfield, C. (2011). Robust averaging during perceptual judgment. *Proceedings of the National Academy of Sciences of the United States of America, 108*(32), 13341–13346. https://doi.org/10.1073/pnas.1104517108

Deza, A., Chen, Y.-C., Long, B., & Konkle, T. (2019). *Accelerated Texforms: Alternative Methods for Generating Unrecognizable Object Images with Preserved Mid-Level Features*. https://doi.org/10.32470/ccn.2019.1412-0

Elias, E., Dyer, M., & Sweeny, T. D. (2017). Ensemble Perception of Dynamic Emotional Groups. *Psychological Science, 28*(2), 193–203. https://doi.org/10.1177/0956797616678188

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is "Special" about Face Perception? *Psychological Review, 105*(3), 482–498. https://doi.org/10.1037/0033-295X.105.3.482

Farah, M. J., Wilson, K. D., Maxwell Drain, H., & Tanaka, J. R. (1995). The inverted face inversion effect in prosopagnosia: Evidence for mandatory, face-specific perceptual mechanisms. *Vision Research, 35*(14), 2089–2093. https://doi.org/10.1016/0042-6989(94)00273-O

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191

Florey, J., Clifford, C. W., Dakin, S., & Mareschal, I. (2016). Spatial limitations in averaging social cues. *Scientific reports, 6*, 32210. https://doi.org/10.1038/srep32210

Florey, J., Dakin, S. C., & Mareschal, I. (2017). Comparing averaging limits for social cues over space and time. *Journal of Vision*, *17*(9). https://doi.org/10.1167/17.9.17

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience, 14*(9), 1195–1204. https://doi.org/10.1038/nn.2889

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*(1714). https://doi.org/10.1098/rstb.2016.0102

Haberman, J., & Suresh, S. (2021). Ensemble size judgments account for size constancy. *Attention, Perception, and Psychophysics, 83*(3), 925–933. https://doi.org/10.3758/s13414-020-02144-6

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology, 17*(17), R751–R753. https://doi.org/10.1016/j.cub.2007.06.039

Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception & Performance, 35*(3), 718–734. https://doi.org/10.1037/a0013899

Haberman, J., & Whitney, D. (2011). Efficient summary statistical representation when change localization fails. *Psychonomic Bulletin & Review*, *18*(5),855–859.https://doi.org/10.3758/s13423-011-0125-6

Han, L., Leib, A. Y., Budish, D., & Whitney, D. (2020). Holistic Ensemble Perception. *Journal of Vision, 19*(10), 194b. https://doi.org/10.1167/19.10.194b

Hayes, T. R., & Henderson, J. M. (2021). Looking for Semantic Similarity: What a Vector-Space Model of Semantics Can Tell Us About Attention in Real-World Scenes. *Psychological Science, 32*(8), 1262–1270. https://doi.org/10.1177/0956797621994768

Huang, H. Y. (2016). Mixture random-effect IRT models for controlling extreme response style on rating scales. *Frontiers in Psychology*, *7*(NOV). https://doi.org/10.3389/fpsyg.2016.01706

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. https://doi.org/10.1113/jphysiol.1959.sp006308

JASP Team. (2021). JASP (Version 0.14.0.0) [Computer software]. Amsterdam, the Netherlands: JASP

Jeffreys, H. (1961). *Theory of probability* (3rd edn.). Oxford: Oxford University Press.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17*(11), 4302–4311. https://doi.org/10.1523/jneurosci.17-11-04302.1997

Khayat, N., Fusi, S., & Hochstein, S. (2021). Perceiving ensemble statistics of novel image sets. *Attention, Perception, and Psychophysics, 83*(3), 1312–1328. https://doi.org/10.3758/s13414-020-02174-0

Khayat, N., & Hochstein, S. (2019). Relating categorization to set summary statistics perception. *Attention, Perception, and Psychophysics*. https://doi.org/10.3758/s13414-019-01792-7

Koenderink, J. J., van Doorn, A. J., & Pont, S. C. (2004). Light direction from shad(ow)ed random gaussian surfaces. *Perception, 33*(12), 1405–1420. https://doi.org/10.1068/p5287

Konkle, T., & Caramazza, A. (2013). Tripartite Organization of the Ventral Stream by Animacy and Object Size. *Journal of Neuroscience, 33*(25), 10235–10242. https://doi.org/10.1523/JNEUROSCI.0983-13.2013

Leib, A. Y., Fischer, J., Liu, Y., Qiu, S., Robertson, L., & Whitney, D. (2014). Ensemble crowd perception : A viewpoint-invariant mechanism to represent average crowd identity. *Journal of Vision, 14*(8), 1–13. https://doi.org/10.1167/14.8.26.doi

Leib, A. Y., Kosovicheva, A., & Whitney, D. (2016). Fast ensemble representations for abstract visual impressions. *Nature Communications, 7*, 13186. https://doi.org/10.1038/ncomms13186

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. *Perception and Psychophysics, 63*(4), 676–697. https://doi.org/10.3758/BF03194429

Li, S. P. D., & Bonner, M. (2020). Curvature as an Organizing Principle of Mid-level Visual Representation: A Semantic-preference Mapping Approach. *NeurIPS 2020 Workshop SVRHM*. https://openreview.net/forum?id=CUi1G2UWsAm%0Ahttps://openreview.net/pdf?id=CUi1G2UWsAm

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

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 of the United States of America, 115*(38), E9015–E9024. https://doi.org/10.1073/pnas.1719616115

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*(6657), 279–281. https://doi.org/10.1038/36846

Mack, A., & Rock, I. (1998). Inattentional blindness. *MIT Press/Bradford Books Series in Cognitive Psychology*, *5*(3), Inattentional blindness. xiv, 273. https://doi.org/10.1016/j.aorn.2010.03.011

Markov, Y. A., & Tiurina, N. A. (2021). Size-distance rescaling in the ensemble representation of range: Study with binocular and monocular cues. *Acta Psychologica*, *213*. https://doi.org/10.1016/j.actpsy.2020.103238

Maule, J., & Franklin, A. (2015). Effects of ensemble complexity and perceptual similarity on rapid averaging of hue. *Journal of Vision, 15*(4), 6. https://doi.org/10.1167/15.4.6

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 101*(2), 343–352. https://doi.org/10.1037/h0043158

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems*, *14*(3), 391–412

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience, 4*(7), 739–744. https://doi.org/10.1038/89532

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology, 70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Peirce, J. W. (2015). Understanding mid-level representations in visual processing. *Journal of Vision*, *15*(7). https://doi.org/10.1167/15.7.5

Perrinet, L. U., & Bednar, J. A. (2015). Edge co-occurrences can account for rapid categorization of natural versus animal images. *Scientific Reports*, *5*. https://doi.org/10.1038/srep11400

Piazza, E. A., Sweeny, T. D., Wessel, D., Silver, M. A., & Whitney, D. (2013). Humans Use Summary Statistics to Perceive Auditory Sequences. *Psychological Science, 24*(8), 1389–1397. https://doi.org/10.1177/0956797612473759

Purcell, D. G., & Stewart, A. L. (2010). Still another confounded face in the crowd. *Attention, Perception, & Psychophysics, 72*(8), 2115–2127. https://doi.org/10.3758/APP.72.8.2115

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision, 3*(3), 179–197

Richler, J. J., Cheung, O. S., & Gauthier, I. (2011). Holistic processing predicts face recognition. *Psychological Science, 22*(4), 464–471. https://doi.org/10.1177/0956797611401753

Robinson, M. M., & Brady, T. F. (2023). A quantitative model of ensemble perception as summed activation in feature space.

Nature Human Behaviour, 7(10), 1638–1651. https://doi.org/10.1038/s41562-023-01602-z

Rosenholtz, R., Huang, J., Raj, A., Balas, B. J., & Ilie, L. (2012). A summary statistic representation in peripheral vision explains visual search. *Journal of Vision, 12*(4), 14. https://doi.org/10.1167/12.4.14

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

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

Simons, D. J., & Chabris, C. F. (1999). Gorillas in Our Midst: Sustained Inattentional Blindness for Dynamic Events. *Perception, 28*(9), 1059–1074. https://doi.org/10.1068/p281059

Stojanoski, B., & Cusack, R. (2014). Time to wave good-bye to phase scrambling: Creating controlled scrambled images using diffeomorphic transformations. *Journal of Vision*, *14*(12). https://doi.org/10.1167/14.12.6

Sweeny, T. D., Bates, A., & Elias, E. (2021). Ensemble perception includes information from multiple spatial scales. *Attention, Perception, and Psychophysics, 83*(3), 982–997. https://doi.org/10.3758/s13414-020-02109-9

Sweeny, T. D., & Whitney, D. (2014). Perceiving Crowd Attention: Ensemble Perception of a Crowd's Gaze. *Psychological Science, 25*(10), 1903–1913. https://doi.org/10.1177/0956797614544510

Sweeny, T. D., Whitney, D., & Haroz, S. (2013). Perceiving group behavior: Sensitive ensemble coding mechanisms for biological motion of human crowds. *Journal of Experimental Psychology: Human Perception and Performance, 39*(2), 329–337. https://doi.org/10.1037/a0028712

Sweeny, T. D., Wurnitsch, N., Gopnik, A., & Whitney, D. (2015). Ensemble perception of size in 4–5-year-old children. *Developmental Science, 18*(4), 556–568. https://doi.org/10.1111/desc.12239

Tiurina, N. A., & Utochkin, I. S. (2019). Ensemble perception in depth: Correct size-distance rescaling of multiple objects before averaging. *Journal of Experimental Psychology: General, 148*(4), 728–738. https://doi.org/10.1037/xge0000485

Utochkin, I. S., Choi, J., & Chong, S. C. (2024). A population response model of ensemble perception. *Psychological review*, *131*(1), 36–57. https://doi.org/10.1037/rev0000426

Vallat, R. (2018). Pingouin: statistics in Python. *Journal of Open Source Software*, *3*(31), 1026. https://doi.org/10.21105/joss.01026

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kesteren, E.-J., van Doorn, J., Šmíra, M., Epskamp, S., Etz, A., Matzke, D., … Morey, R. D. (2017). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, 1–19. https://doi.org/10.3758/s13423-017-1323-7

Wang, R., Janini, D., & Konkle, T. (2022). Mid-level Feature Differences Support Early Animacy and Object Size Distinctions: Evidence from Electroencephalography Decoding. *Journal of Cognitive Neuroscience, 34*, 1670–1680

Watamaniuk, S. N. J., & McKee, S. P. (1998). Simultaneous encoding of direction at a local and global scale. *Perception and Psychophysics, 60*(2), 191–200. https://doi.org/10.3758/BF03206028

Watamaniuk, S. N. J., Sekuler, R., & Williams, D. W. (1989). Direction perception in complex dynamic displays: The integration of direction information. *Vision Research, 29*(1), 47–59. https://doi.org/10.1016/0042-6989(89)90173-9

Whiting, B. F., & Oriet, C. (2011). Rapid averaging? Not so fast! *Psychonomic Bulletin & Review, 18*(3), 484–489. https://doi.org/10.3758/s13423-011-0071-3

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble Perception. *Annual Review of Psychology, 69*, 105–129. https://doi.org/10.1146/annurev-psych-010416-044232

Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Gosselin, F., & Tanaka, J. W. (2010). Controlling low-level image properties: The SHINE toolbox. *Behavior Research Methods, 42*(3), 671–684. https://doi.org/10.3758/BRM.42.3.671

Wolfe, B. A., Kosovicheva, A. A., Wood, K., & Whitney, D. (2015). Foveal input is not required for perception of crowd facial expression. *Journal of Vision*, *15*(4). https://doi.org/10.1167/15.4.11

Yamanashi Leib, A., Chang, K., Xia, Y., Peng, A., & Whitney, D. (2020). Fleeting impressions of economic value via summary statistical representations. *Journal of Experimental Psychology: General, 149*(10), 1811–1822. https://doi.org/10.1037/xge0000745

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), 1–12. https://doi.org/10.1167/18.12.3