Concept Learning of Ecological and Artificial Stimuli in Rhesus Macaques

Drew Altschul[1,2], **Greg Jensen**[3], **and Herbert S. Terrace**[3]

[1]**University of Edinburgh**
[2]**Scottish Primate Research Group**
[3]**Columbia University**

ABSTRACT

The study of concepts in animals is complicated by the possibility that performance reflects reinforcement learning of discriminative cues, which might be used to categorize of stimuli. To minimize that possibility, we trained seven rhesus macaques to respond, in a specific order, to four simultaneously presented exemplars of different perceptual concepts. These exemplars were drawn at random from large banks of images; in some conditions, the stimuli changed on every trial. Subjects nevertheless identified and ordered these stimuli correctly. Three subjects learned to correctly order ecologically relevant concepts; four subjects, to order close-up sections of paintings by four artists with distinctive styles. All subjects classified stimuli significantly better than that predicted by chance, and outperformed a feature-based computer vision algorithm, even when the exemplars were changed on every trial. Furthermore, six subjects (three using ecological stimuli and three using paintings) transferred these concepts to novel stimuli. Our results suggest that monkeys possess a flexible ability to form class-based perceptual concepts that cannot be explained as the mere discrimination of physical features.

Keywords: cognition, concepts, categories, simultaneous chain, rhesus macaques

INTRODUCTION

Over the last 50 years, considerable effort has been devoted to investigating how non-human animals (hereafter, "animals") perceive and categorize visual stimuli. The resulting literature has demonstrated that animals have the ability to classify a bewildering range of stimuli, including organic forms, such as faces (Marsh and MacDonald, 2008), plants, and animals (Roberts and Mazmanian, 1996; Vonk and MacDonald, 2002; Vonk, 2013), as well as man-made objects, such as cars, chairs (Bhatt et al., 1988), orthographic characters (Schrier et al., 1984), paintings (Watanabe, 2013), cartoons (Matsukawa et al., 2001), and abstract forms (Vogels, 1999). Animals have also correctly identified never-before-seen exemplars, showing that this ability is not limited to experiences with specific stimuli (Schrier and Brady, 1987; Sigala, 2009). These sophisticated abilities have been reviewed extensively elsewhere (Jitsumori and Delius, 2001; Miller et al., 2003; Katz et al., 2007; Zentall et al., 2008).

The interpretations of such findings have stirred controversy because they challenge long-standing assumptions about the nature of *concepts*, as distinguished from categories. Some authors have defined concepts as necessarily linguistic *a priori*, ruling out mechanisms for non-linguistic concept formation (Chater and Heyes, 1994). Others have argued that animals learn to categorize using reinforcement learning and associative conditioning (Roberts, 1996). Under this hypothesis, an animal's classification of similar stimuli (as demonstrated by Herrnstein et al., 1976) relies on associative strength of discriminable features common across a category (Lea, 1984).

Herrnstein (1990) attempted to integrate linguistic and associative accounts by proposing that animals

use *open-ended categories* to classify stimuli, rather than concepts. The criteria for this distinction were vague: a discrimination could be attributed to a concept only if a characteristic other than similarity was used to classify novel exemplars. The inevitable skeptical retort argued that, because stimuli must necessarily have some features in common (without which they would be indistinguishable), those features must permit categorization on the basis of their similarity (Huber, 2000). Thus, Herrnstein's proposal leaves the issue unresolved, and enduring skepticism about concept learning in animals has assumed that feature-based associations can explain performance.

Despite this skepticism, animals have nevertheless demonstrated sophisticated categorization aptitudes. For example, monkeys can discriminate ecologically relevant objects (e.g. animals) from man-made objects (e.g. umbrellas) (Bhatt et al., 1988; Crouzet et al., 2012). Moreover, this aptitude was not specific to stimuli that may have previously had evolutionary significance. Manufactured objects were identified with the same accuracy and speed as natural scenes and organisms (Roberts and Mazmanian, 1996). These findings point to a highly general learning mechanism that can be used to learn discriminations never found in an organism's evolutionary niche.

The persistence of the controversy has been complicated by two methodological problems, described by Jensen and Altschul (2015). One is that subjects may develop a *tailor-made classifier* during training, which solves the discrimination using features as shortcuts. For example, discriminating between faces and houses may require only attending to the squareness of windows vs. the roundness of eyes. Post-hoc analysis cannot rule out non-conceptual feature-based strategies. However, as the number of categories is increased, learning a correspondingly complicated list of rules becomes an inefficient strategy. A second difficulty is that binary discrimination tests allow for *educated guesses* even when understanding falls short of conceptual. When training and testing rely on dichotomous discriminations (e.g. Roberts and Mazmanian, 1996; Vogels, 1999) performance can exceed chance using simple feature-based strategies. More difficult test procedures can, however, dramatically reduce the false-positive rate.

A noteworthy exception to these design problems is a study reported by Bhatt et al. (1988). In a series of experiments, pigeons learned to discriminate between four varieties of stimulus (e.g. photographs of cats, flowers, cars, and chairs). Because these four conceptual groupings were trained in parallel, they protected against tailor-made classifiers. Furthermore, selecting from four alternatives at test lowered the false-positive rate of guessing to 25% (from the more typical 50%). To date, it serves as a landmark study, and a high water mark for evidentiary rigor. But while this study is methodologically excellent, subjects' performance could still have been based on consistent image features (eye, wheels, petals, etc.).

In an effort to provide more compelling evidence of an animal's conceptual abilities, we trained subjects to classify stimuli during a cognitively demanding task. Instead of training concepts one at a time, we used a variation of the simultaneous chaining paradigm (or "SimChain"; Terrace, 1984) to train four different stimulus categories simultaneously. Following training, subjects had to categorize four stimuli that were randomly drawn from large banks of exemplars. In addition to training four concepts at once, SimChain also provided a strenuous classification test, because the odds of completing a trial successfully by chance alone are less than 5%. Rapid, accurate performance on this task would constrain claims made by both proponents and skeptics of animal concept formation.

In Experiment 1, three rhesus monkeys were trained to classify four naturally occurring concepts: birds, flowers, cats, and people. Exemplars were randomly drawn from large banks of photographs for every trial. Subjects were then required to select exemplars (one for each concept) in a prescribed order. In Experiment 2, four macaques were trained to classify cropped close-ups from works of four painters: Claude Monet, Vincent van Gogh, Salvador Dalí, and Jean-Léon Gérôme. These artificial stimuli had no discrete features (e.g. faces, wings) in common. Accordingly, subjects had to attend to the gestalt properties that distinguished each painterly style. The logic of this approach allowed us to assess a

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.967v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

subject's ability to classify exemplars of different concepts with high accuracy and fast reaction times under conditions that rendered brute force strategies (e.g. memorization or discrete feature analysis) inordinately costly and unreliable.

EXPERIMENT 1: PHOTOGRAPHS OF ECOLOGICAL CATEGORIES

Methods

Subjects: Data were collected from 3 male rhesus monkeys (*Macaca mulatta*), Augustus, Coltrane, and Lashley. All subjects had extensive experience with the standard simultaneous chain task prior to the experiment. Subjects were housed at the New York State Psychiatric Institute throughout the study. Treatment conformed with the guidelines set by the U.S. Department of Health and Human Services (National Institute of Health) for the care and use of laboratory animals. The study was approved under protocol AC-AAAB1238 by the Institutional Animal Care and Use Committee at Columbia University. In addition to pellets obtained during experimentation, subjects were given a mixed diet of primate chow and fruit immediately following daily testing. Water was available *ad libitum*.

Apparatus: Subjects performed tasks in operant chambers made of Plexiglas and stainless steel (53 cm $\times$ 48cm $\times$ 53 cm) that were enclosed within sound-attenuated booths (127 cm high $\times$ 97 cm wide $\times$ 97 cm deep). Each booth contained a pellet dispenser (Med Associates) that delivered 190-mg banana pellets (Bioserv) and a closed-circuit camera. Subjects responded by touching stimuli that were presented on a touch-sensitive 15-inch (38 cm) computer monitor in the chamber. All experimental tasks were programmed using Real Studio (formerly RealBASIC) and were controlled by an iMac computer (model: MA710xx/A).

Procedure: Subjects performed two closely related tasks. One was the simultaneous chaining task (Terrace, 1984; Jensen et al., 2013a), hereafter identified as the "SimChain" task. The other was the newly designed "Concept Chain" task. Sample trials for each task are depicted in Figure 1A.

In the SimChain task, four photographs were simultaneously presented on the screen. The positions of the stimuli were scrambled from one trial to the next, but the same set of images was used throughout a session. A reward was delivered after every item was touched in the correct order. However, any errors (responses that deviated from the prescribed order) resulted in a 6 second time-out. Because no cues indicated the next item to which an animal should respond, trial and error was the only way to learn the order of an entirely unfamiliar set of stimuli. Examples of a correct trial and two incorrect trials are depicted in Figure 1B. Sessions lasted 40 trials.

The Concept Chain task closely resembled the SimChain task in that subjects could earn rewards only if they touched four stimuli in a prescribed order. Unlike a SimChain, however, some or all of the stimuli in a Concept Chain changed randomly after every trial. For example, when the 'flowers' stimulus was set to vary, a different flower was presented on each trial, drawn at random from a large image bank. Concept Chain sessions also lasted 40 trials.

In Experiment 1, our stimulus sets consisted of 2867 pictures of people, 3032 pictures of flowers, 872 pictures of cats, and 3110 pictures of birds. Two exemplars from each set are shown in Figure 1C (for others, see Figure **??**). Our criteria for "ecologically valid" stimuli were broad: Photographs taken under reasonably normal lighting conditions, depicting primarily organic content. For example, a picture of a man wearing a shirt is considered 'ecologically valid' because it depicts a man, despite the artificial origin of his shirt. The appendix provides a systematic analysis of low-level properties of the image sets.

Each subject was assigned a different "correct sequence" (Augustus: Flowers $\rightarrow$ Cats $\rightarrow$ People $\rightarrow$ Birds; Coltrane: Birds $\rightarrow$ Flowers $\rightarrow$ Cats $\rightarrow$ People; Lashley: Cats $\rightarrow$ Birds $\rightarrow$ People $\rightarrow$ Flowers). This order was maintained throughout the experiment.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.967v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

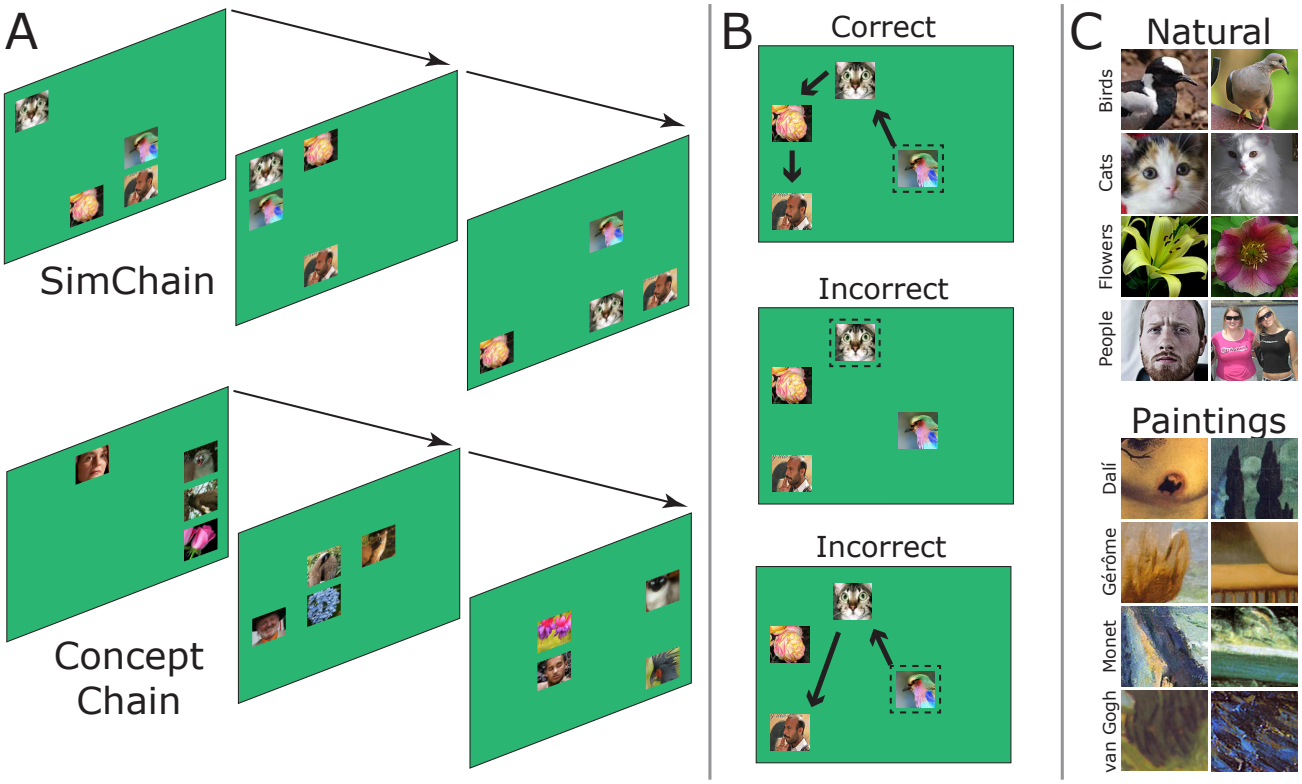

Figure 1. Procedures and stimuli used in Experiments 1 and 2. (A) Three trials of the SimChain and Concept Chain tasks. The Simchain task scrambled stimulus positions on a trial-by-trial basis. The Concept Chain task scrambled positions, but also selected new stimuli at random for some or all categories in every trial. (B) Examples of one correct and two incorrect trials. Dashed boxes indicate the first stimulus touched by the animal, and the arrows show the subsequent touches. The top example was "correct" because all items were touched in the prescribed order, and that trial ended in reward. The other two trials are "incorrect" trials that resulted in a 6 s timeout. The middle example depicts an incorrect initial touch, ending the trial immediately. The bottom example shows two correct touches, followed by an erroneous touch to the fourth list item. Note that an incorrect trial only ends at the first incorrect touch, permitting animals to progress by trial and error. (C) Examples of 'ecologically valid' stimuli used in Experiment 1 and the 'painting' stimuli used in Experiment 2.

The Concept Chain task was introduced in stages. Subjects advanced to the next stage only after they responded correctly to each four items in a session with 80% accuracy. Initially, only one stimulus varied. In the first stage, the 4th item varied randomly, while the other three list items remained the same. In the second stage, the 3rd category's stimulus varied rather than the fourth. Subsequent stages varied the 2nd item, then the 1st item.

Subjects were then trained with two concept stimuli varying: The 3rd and 4th items varied first, followed by the 2nd and 4th, and so on (1st and 4th; 2nd and 3rd; 1st and 3rd; 1st and 2nd). Next, three items were changed every trial (2nd, 3rd, and 4th; 1st, 3rd, and 4th; 1st, 2nd, and 4th; 1st, 2nd, and 3rd). During the final stage, all stimuli changed on every trial. Once the 80% criterion was satisfied, an additional 25 sessions of data were collected, again with all items changing. These 25 sessions are reported in the results as the "Concept Chain" sessions.

Once the Concept Chain sessions were complete, subjects performed two version of the SimChain task. In each of 25 sessions of the "Category SimChain" sessions, stimulus lists consisted of four fixed stimuli that belonged to the learned categories but where novel to subjects. Performance was compared to 35 sessions of "Arbitrary SimChain," which used stimuli unrelated to the trained classifications.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.967v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

Results

Each subject provided evidence that it could discriminate novel exemplars of four simultaneously presented concepts. Performance (accuracy and progress on each lists) and reaction times (RTs) were compared for three conditions: the Concept Chain task (using categorical exemplars that changed randomly on each trial), the Category SimChain task (using a fixed list of four categorical exemplars), and the Arbitrary SimChain task (using a fixed set of arbitrary stimuli).

Performance in both tasks was characterized using a learning curve originally identified by Thurstone (1919):

$$y = \frac{L\,(x+P)}{x+P+R} \tag{1}$$

Here, 'average number of correct responses in a list' y is predicted in terms of trial x and three parameters: L (the maximum possible level of performance), R (the 'learning cost,' such that lower values correspond to faster learning), and P (the prior knowledge, scaled according to R). In a 4-item SimChain, the value of L is defined by the task to be 4.

Thurstone's learning curve describes a diminishing-returns growth function. The learning cost R governs curvature, such that each interval doubles the time it takes to make half the progress towards the maximum L. It takes R trials to progress from $y = 0.0$ to $0.5L$, but it takes an additional $2R$ trials to subsequently progress from $y = 0.5L$ to $0.75L$.

The P parameter shifts the curve horizontally, allowing subjects with prior knowledge to exceed chance performance on the first trial. Because the value of P is set relative to R, we use the standardized metric $\frac{P}{R}$. When subjects are naïve, $\frac{P}{R}$ is expected to equal 0.0, whereas subjects bringing prior knowledge to the task have higher values of $\frac{P}{R}$, with no upper limit to its value.

This learning curve provides a description of overall performance, and is not a prediction of performance in an individual session. Trial-by-trial progress in SimChain performance is often abrupt, as a function of whether subjects make lucky or unlucky guesses (Jensen, 2013). The learning curves presented here capture averages of performance across many lists, and therefore show a general aptitude for list learning.

In order to estimate model parameters, the following linearization is used:

$$\left(\frac{L}{y} - 1\right)^{-1} = \frac{x}{R} + \frac{P}{R} \tag{2}$$

Because Equation 2 does not have uniform residual variance, bootstrapping was used to obtain the sample variance for each transformed dependent average, prior to a weighted regression, as described by Jensen et al. (2013a).

Figure 2 shows performance for each subject in the Concept Chain task. The observed flat slopes (i.e. very large R) and high $\frac{P}{R}$ ratios (consistently greater than 1) reflect ceiling levels of performance. Given their successful completion of the training stages, the first trial in a session should be no different from the last because each trial presents a new combination of stimuli. Performance substantially exceeded chance levels, demonstrating proficient ordering of dynamic stimuli in an order prescribed by their category membership.

Figure 3 compares performance on the two types of SimChain tasks following training. Category SimChain performance (blue circles) consistently exceeded Arbitrary SimChain performance (green diamonds) for all subjects. The confidence intervals for the regression parameters confirm these differences, which manifested in several ways. According to post-hoc tests for the individual regression parameters,

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.967v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

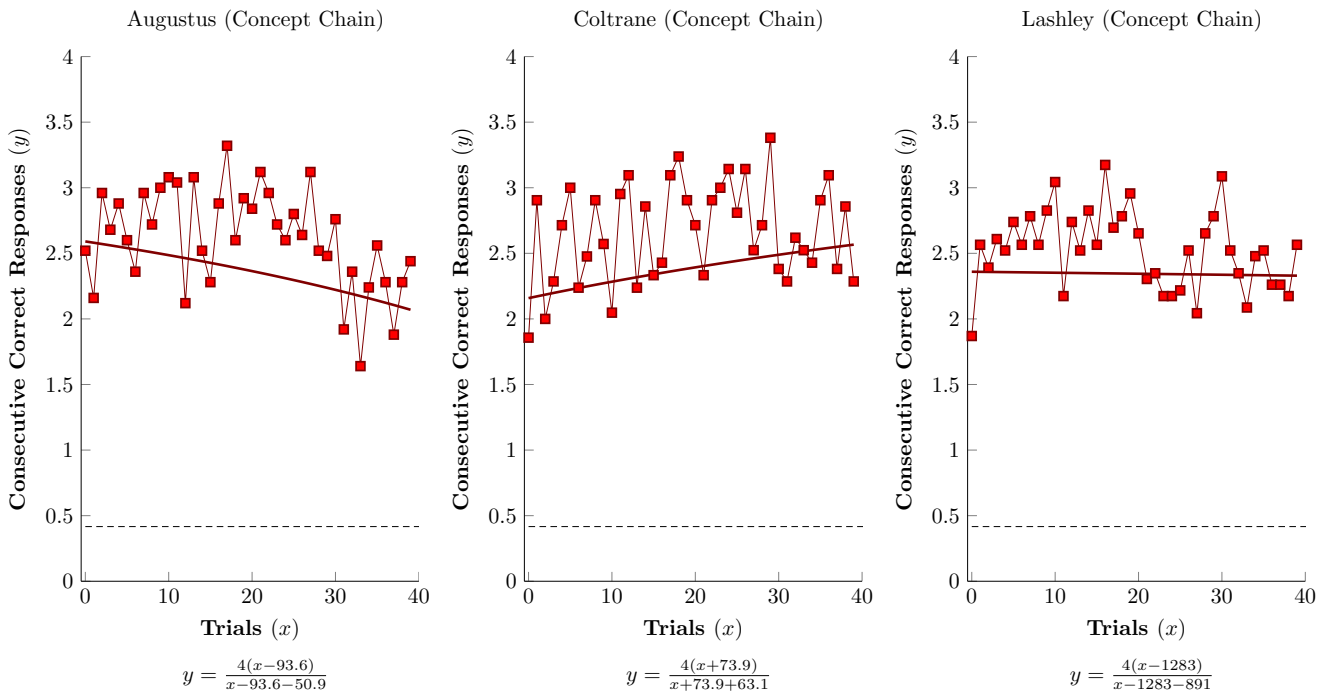

Figure 2. Performance during the Concept Chain task in Experiment 1. Points represent trial-by-trial averages of 25 sessions, whereas the heavy lines represent the model fit of Equation 1 (parameters below each plot). The horizontal dashed line shows chance performance.

Augustus had a significantly lower R parameter ($t(76) > 6.12, p < .001$) and a significantly higher $\frac{P}{R}$ ($t(76) > 3.65, p < .001$), both indicating better performance. Coltrane's R parameter was significantly lower ($t(76) > 4.68, p < .001$), but $\frac{P}{R}$ did not differ significantly ($t(76) = 1.96, p = .053$). Finally, Lashley had a higher $\frac{P}{R}$ ($t(76) > 6.35, p < .001$), but R did not differ significantly ($t(76) = 0.45, p = .65$).

Although it is straightforward to demonstrate that subjects exceeded chance levels of accuracy, it is considerably more difficult to specify the performance expected under the null hypothesis that "subjects responded only on the basis of feature associations." Behavior driven only by feature-based associations would certainly exceed chance, but it is unclear by how much. We therefore drew upon the machine learning literature and used the "bag-of-features" image classifier (O'Hara and Draper, 2011) as a candidate for this null hypothesis. The bag-of-features classifier is a sophisticated algorithm, but its discriminations are ultimately only made on the basis of low-level statistical regularities. This makes it a reasonable stand-in for a strictly associative model. Insofar as subjects outperformed the algorithms, we take this as evidence that subjects' strategy was more than merely feature-based.

Like subjects, the bag-of-features classifier was trained on all four categories of images simultaneously. Half of the images were used as a training set, and classifier performance was then validated using the other half (as described by Jensen and Altschul, 2015). This yielded the following "confusion matrix,"

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.967v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

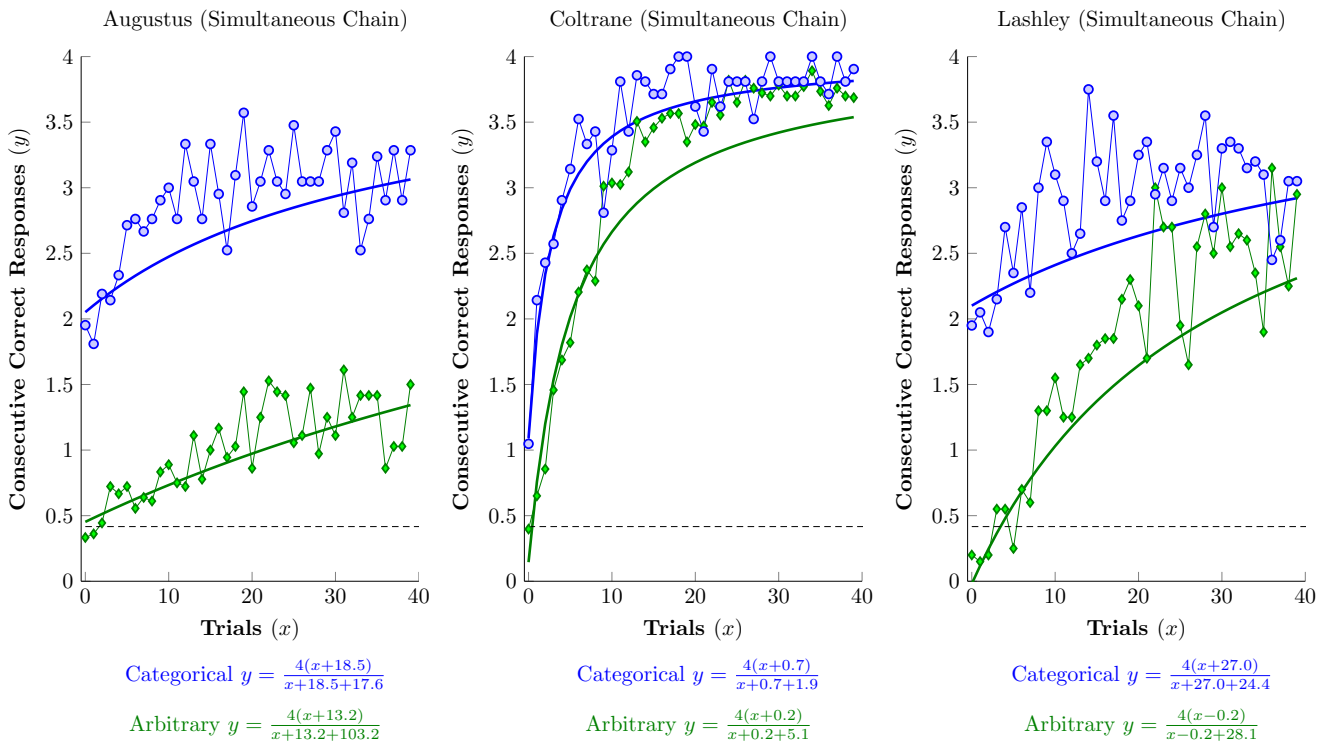

Figure 3. Performance during SimChain tasks in Experiment 1, given categorical stimuli (blue circles) or arbitrary stimuli (green diamonds). Points represent trial-by-trial averages of 25 categorical and 35 arbitrary sessions. The heavy curved lines represent the model fit of Equation 1 (parameters below each plot). The horizontal dashed line shows chance performance.

which provides the odds of correctly categorizing any single stimulus:

$$\begin{array}{cc} & \text{Guessed Category} \\ & \begin{array}{cccc} \text{birds} & \text{cats} & \text{flowers} & \text{people} \end{array} \\ \text{True Category} \begin{array}{c} \text{birds} \\ \text{cats} \\ \text{flowers} \\ \text{people} \end{array} & \begin{bmatrix} 0.64 & 0.10 & 0.11 & 0.15 \\ 0.13 & 0.62 & 0.15 & 0.10 \\ 0.11 & 0.15 & 0.62 & 0.11 \\ 0.11 & 0.10 & 0.14 & 0.66 \end{bmatrix} \end{array} \tag{3}$$

Thus, if the algorithm was presented with a photo of a bird, there was a 0.64 probability of correctly identifying it as a bird stimulus, and a 0.15 probability that it would be misidentified as a photo of a person. This constitutes a moderate level of identifiably for such an algorithm: Performance is well above chance, but there were nonetheless many errors.

In order to perform the SimChain task, the algorithm had to identify all four stimuli simultaneously, which it had a probability of 0.16 of doing perfectly. However, if more than one stimulus was identified as belonging to a given category, the algorithm guessed and, if the guess was made correctly, continued on the remaining stimuli by process of elimination. When this kind of guessing was taken into account, the algorithm was able to complete the SimChain and earn a reward on approximately 36% of trials (depending on the order of the categories).

Figure 4 shows the conditional probability of responding correctly to the nth item in a list (with 95% confidence intervals) for Category SimChain trials (blue circles), Concept Chain trials (red squares), and Arbitrary SimChain trials (green diamonds). Also shown is the simulated performance of the bag-of-features algorithm (black squares), given the confusion matrix presented in Eq. 3 and the stimulus order

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.967v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

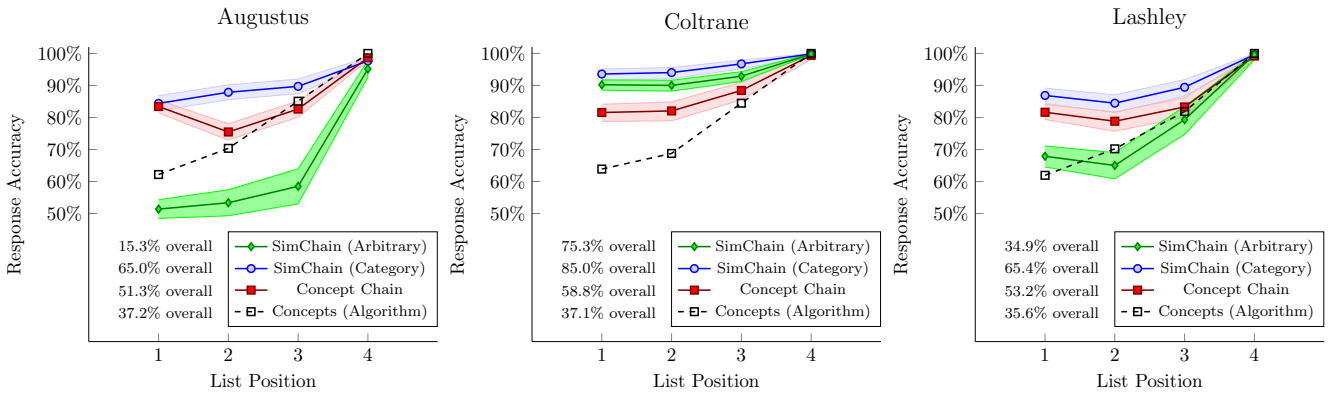

Figure 4. Conditional probability of correct response to items at each list position for each subject in Experiment 1. Plotted is response accuracy for Category SimChain trials (blue circles), Concept Chain trials (red squares), and Arbitrary SimChain trials (green diamonds). Lines connect averages that are conditional upon one another. Overall accuracy (next to the legend entry for each task type) indicates a subject's probability of successfully earning a reward at the end of a trial. Each line's shaded region represents the 95% confidence interval for the mean.

experienced by that animal. The legend of Figure 4 also indicates the proportion of trials resulting in a reward in each case. For example, although Augustus had a 0.153 probability of completing an Arbitrary SimChain trial, his probability of making a correct response to the last item exceeded 0.9, conditional on correct responses to the first three items. For each subject, each overall proportion correct was significantly different from all others, according to an omnibus chi-square test (Augustus: $\chi^2 > 568, df = 2, p < .001$; Coltrane: $\chi^2 > 152, df = 2, p < .001$; Lashley $\chi^2 > 151, df = 2, p < .001$) and post-hoc pairwise chi-square tests ($p < .001$) corrected for multiple comparisons using the Holm-Šidák procedure. All subjects also earned rewards on a significantly greater proportion of trials than did the bag-of-features algorithm in both the Concept Chain and the Category SimChain trials, according to a binomial test ($p < .001$).

Figure 5 shows the mean of subjects' log-scaled reaction times (RTs) of responses at each position in the chain (+/- 1 standard error). The sum of these four time intervals constitutes the time needed to complete one trial. RTs accelerated as subjects progressed through the list, consistent with the process-of-elimination character of the SimChain task. Every subject showed a significant difference in RT as a function of list position (Augustus: $F(3, 9248) > 1136, p < .001$; Coltrane: $F(3, 10421) > 1366, p < .001$; Lashley: $F(3, 7327) > 2281, p < .001$). The effect size of these differences was substantial (Augustus: $\omega^2 = 0.244$; Coltrane: $\omega^2 = 0.273$; Lashley: $\omega^2 = 0.460$), dominating the proportion of variance explained. Although significant differences for task type were observed (Augustus: $F(2, 9248) > 123, p < .001$; Coltrane: $F(2, 10421) > 18.4, p < .001$; Lashley: $F(2, 7327) > 79.3, p < .001$), the effect sizes for these differences were negligible (Augustus: $\omega^2 = 0.018$; Coltrane: $\omega^2 = 0.001$; Lashley: $\omega^2 = 0.011$). Similarly, although a significant interaction between these two main effects was observed (Augustus: $F(6, 9248) > 41.2, p < .001$; Coltrane: $F(6, 10421) > 19.3, p < .001$; Lashley: $F(6, 7327) > 19.7, p < .001$), the corresponding effect sizes were very small (Augustus: $\omega^2 = 0.017$; Coltrane: $\omega^2 = 0.004$; Lashley: $\omega^2 = 0.008$).

All three subjects performed well above chance on the Concept Chain task, and also performed better on Categorical SimChain lists than Arbitrary SimChain lists. These results show that monkeys are capable of simultaneously distinguishing between four distinct ecological concepts. Additionally, reaction times were similar in each of the tasks, suggesting that recognizing exemplars of familiar perceptual concepts was no more time-consuming than recognizing familiar items, even when the stimuli changed for every

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.967v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

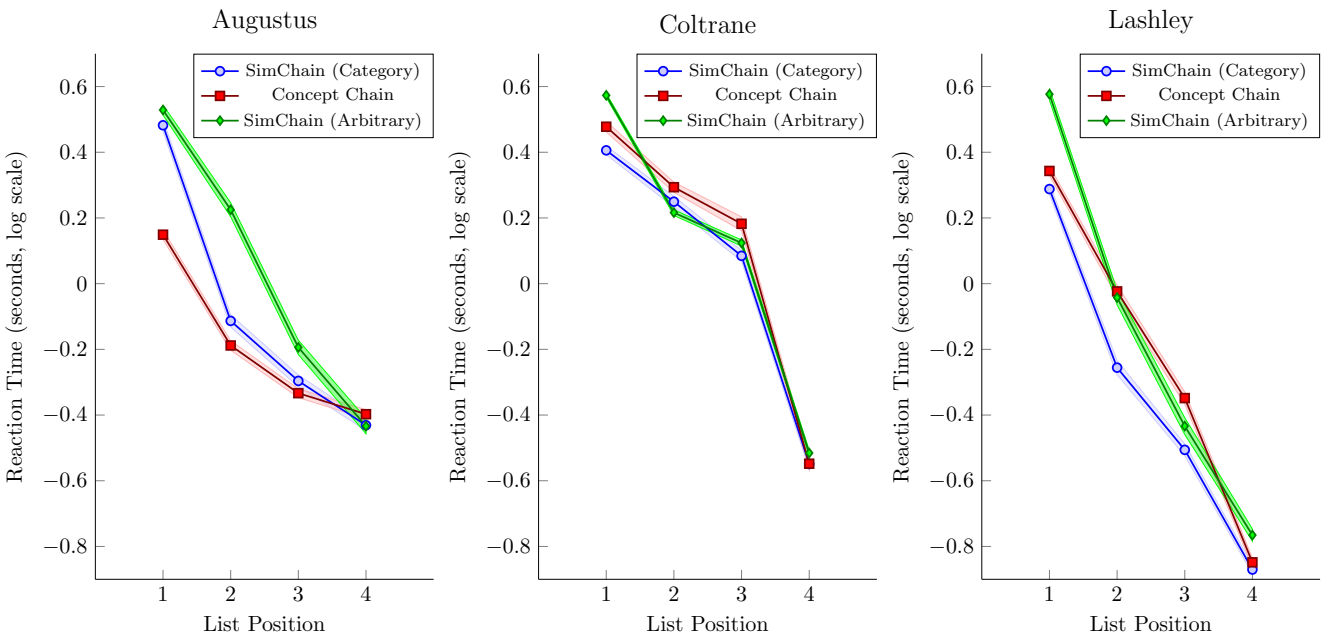

Figure 5. Average of log-scaled reaction times conditional on progress in each trial of Experiment 1. For example, List Position 1 times show to interval prior to the first touch, whereas List Position 2 times correspond to reaction times given that the first touch was correct. Plotted is response accuracy for Category SimChain trials (blue circles), Concept Chain trials (red squares), and Arbitrary SimChain trials (green diamonds), with shaded regions depicting one standard error. Included are 9254 RTs from Augustus, 10427 RTs from Coltrane, and 7333 RTs from Lashley.

trial.

Subjects' conceptual proficiency was impressive given the diversity of stimuli. The category 'cats' not only included housecats but also the larger wild species, such as tigers. 'People' included close-ups of human faces in portrait and profile, as well as photos of crowds. Although most stimuli in each set were in color, black-and-white stimuli were also included. Consequently, no single feature-based rule could be relied upon as a shortcut to execute the sequences required by the Concept Chain task.

Nevertheless, statistical regularities existed within each image set, and these might have been exploited as discriminative cues. As described in the appendix, each set had a distinct distribution of image hue and saturation. To evaluate whether the subject performance and RTs necessarily depended on these image properties, we performed a second experiment using samples from the works of four prolific painters. The resulting sets of images could not be distinguished on the basis of either their discrete features (which varied too much within set to provide usable cues), or on the basis of their low-level image characteristics (which were sufficiently similar to prevent straightforward classification).

EXPERIMENT 2: ABSTRACT STIMULI FROM CANVAS PAINTINGS

Methods

Subjects & Apparatus: Subjects were 4 male rhesus monkeys, Benedict, Horatio, Macduff, and Prospero. These subjects were housed under identical conditions, and were trained using the same apparatus as those in Experiment 1. Although subjects were familiar with the SimChain task from previous experiments, they were otherwise naïve.

Procedure: The SimChain and Concept Chain tasks were employed with four new categories of stimulus: Exemplars were samples from painted artworks. High-resolution images were obtained of

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.967v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

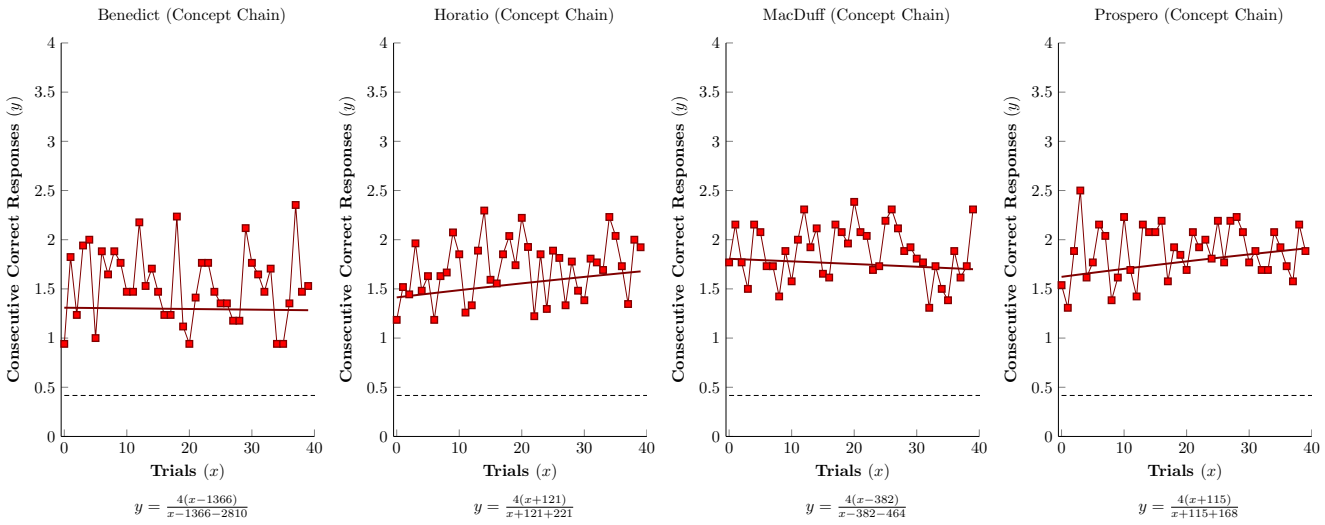

Figure 6. Performance during the Concept Chain task in Experiment 2. Points represent trial-by-trial averages of 25 sessions. The heavy curved lines represent the model fit of Equation 2 (parameters below each plot). The horizontal dashed line shows chance performance.

paintings by four artists with very different styles of painting (Jean-Léon Gérôme, Vincent van Gogh, Claude Monet, and Salvador Dalí). These were sampled to produce close-up exemplars, each 140px by 130px in size (see Figures 1C and 10). The size of the stimulus sets was 203 for Dalí, 406 for Gérôme, 283 for Monet, and 232 for van Gogh.

As in Experiment 1, each subject had a different prescribed order (Benedict: Dalí → Gérôme → Monet → van Gogh; Horatio: van Gogh → Monet → Gérôme → Dalí; MacDuff: Monet → Dalí→ van Gogh → Gérôme; Prospero: Gérôme → van Gogh → Dalí → Monet).

Stimuli could be classified on the basis of brush stroke, texture, contour sharpness, and other aspects of style. Lacking any training in art history, the monkeys needed to become sensitive to these gestalt properties in order to make correct classifications. An analysis of low-level image properties (provided in the supplemental information) confirms that the stimulus sets in Experiment 2 were more internally diverse than those in Experiment 1. This analysis also shows that the four sets were more difficult to discriminate on the basis of simple image statistics alone.

Training proceeded as in Experiment 1, except that additional training stages were added. During the first stage of training, in which a single stimulus varied, the changing stimulus was first selected at random from two images in the stimulus set, then three, five, ten, twenty-five, fifty, and finally a hundred. Subjects advanced through these "set size" stages whenever they satisfied a 70% accuracy criterion for each item. Once subjects reached the stage at which two items were allowed to change, training proceeded exactly as in Experiment 1. Subjects culminated their Concept Chain training with 25 sessions in which every stimulus varied on every trial. These were then followed by 25 sessions of Categorical SimChain using novel painting stimuli, and 35 sessions of the Arbitrary SimChain task.

Results

As in Experiment 1, SimChain performance was modeled using Thurstone's learning curve (Equation 1) and its linearization (Equation 2). Again, the variance of each transformed value y was estimated by bootstrapping in order to allow parameters to be fit using a weighted least squares regression.

Figure 6 shows Concept Chain performance as a function of trials for each subject, averaged over the 25 final sessions of the task. Performance consistently exceeded chance levels, but subjects also made

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.967v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

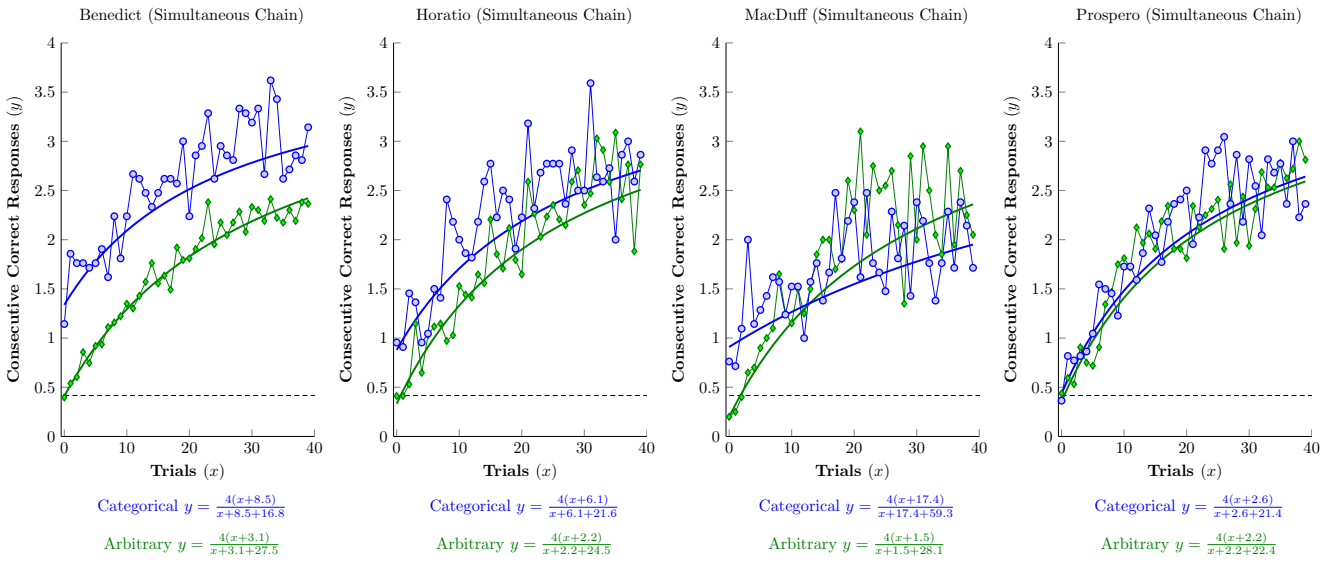

Figure 7. Performance during SimChain tasks in Experiment 2, given categorical stimuli (blue circles) or arbitrary stimuli (green diamonds). Points represent trial-by-trial averages of 25 categorical and 35 arbitrary sessions, whereas the heavy curved lines represent the model fit of Equation 1 (parameters below each plot). The horizontal dashed line shows chance performance.

more errors than in Experiment 1. This was consistent with the intuition that the stimuli in Experiment 2 were considerably more difficult to classify than those in Experiment 1.

Figure 7 compares Category SimChain performance (blue circles) to Arbitrary SimChain performance (green diamonds). Two subjects (Benedict and Horatio) benefitted from prior concept learning, while a third (MacDuff) showed elevated initial responding but a shallower slope. The fourth subject (Prospero) did not appear to use the concept information in these static SimChains, despite having performance comparable to other subjects during training.

A comparison of model residuals confirmed that concept training helped improved Benedict and Horatio's performance. According to post-hoc tests of regression parameters, Benedict had a significantly lower R parameter ($t(76) > 6.96, p < .001$) and a significantly higher $\frac{P}{R}$ ($t(76) > 4.14, p < .001$). Horatio had a significantly higher $\frac{P}{R}$ ($t(76) > 3.40, p < .001$), but his R parameter was not significantly different ($t(76) = 0.89, p = .38$). MacDuff's parameters benefitted only partially: a significantly higher $\frac{P}{R}$ ($t(76) > 5.46, p < .001$) but also a significantly higher (i.e. less efficient) R parameter ($t(76) > 4.80, p < .001$). Finally, Prospero showed no difference in either parameter ($t(76) > 0.48, p > .40$).

As in Experiment 1, the bag-of-features algorithm was trained to classify the stimuli from the four painters. This analysis revealed that although the algorithm was effective at identifying Monet, it did poorly with Dalí and van Gogh (performing below chance in the latter case), as shown in Equation 4:

$$\begin{array}{c} \\ \text{True Category} \end{array} \begin{array}{c} \text{Guessed Category} \\ \begin{array}{cccc} \text{Dalí} & \text{Gérôme} & \text{Monet} & \text{van Gogh} \end{array} \\ \begin{array}{c} \text{Dalí} \\ \text{Gérôme} \\ \text{Monet} \\ \text{van Gogh} \end{array} \begin{bmatrix} 0.46 & 0.36 & 0.07 & 0.11 \\ 0.17 & 0.61 & 0.14 & 0.07 \\ 0.04 & 0.04 & 0.86 & 0.07 \\ 0.16 & 0.19 & 0.43 & 0.22 \end{bmatrix} \end{array} \tag{4}$$

Because the difficulty of identification was so different from one artist to the next, the prescribed order of the categories mattered a great deal in terms of simulating performance.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.967v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

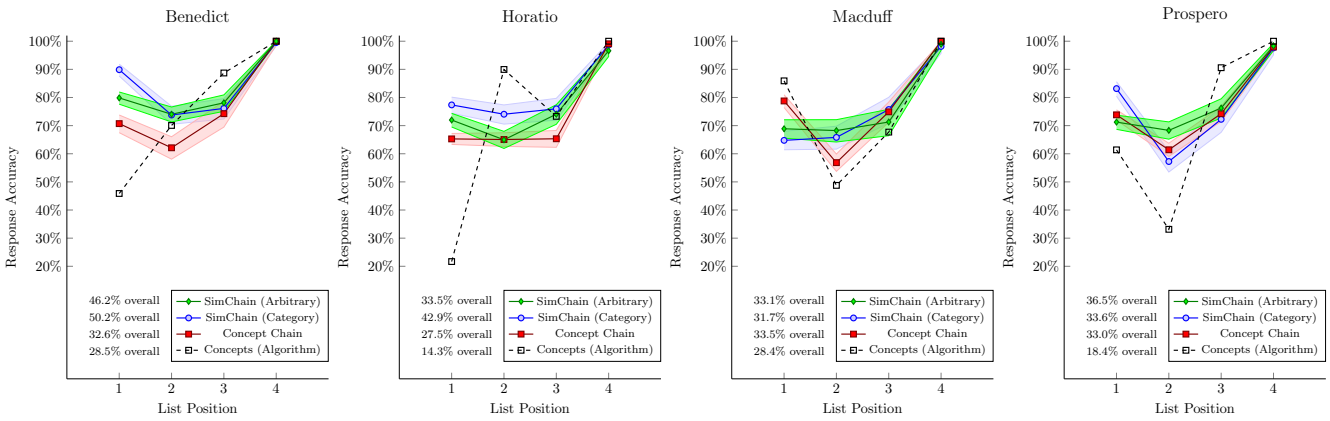

Figure 8. Conditional probability of correct response to items at each list position for each subject in Experiment 2. Plotted is response accuracy for Category SimChain trials (blue circles), Concept Chain trials (red squares), and Arbitrary SimChain trials (green diamonds). Lines connect averages that are conditional upon one another. Overall accuracy (next to the legend entry for each task type) indicates a subject's probability of successfully earning a reward at the end of a trial. The shaded regions around each line correspond to the 95% confidence interval for the mean.

Figure 8 shows conditional probabilities of responding correctly to the nth item in a list (with 95% confidence intervals) for categorical SimChain trials (blue circles), Concept Chain trials (red squares), and arbitrary SimChain trials (green diamonds). The simulated performance of the bag-of-features algorithm is shown in black, and the overall proportion of trials resulting in a reward for each condition is shown next to the legend. The effects of the differential difficulty of the artists for the bag-of-features algorithm is highly evident in these graphs. For example, Horatio's first required response was to Van Gogh (the most difficult), followed by Monet (the least difficult), leading to a dramatic fluctuation in the algorithm's conditional probabilities. Consequently, the algorithm was rewarded anywhere from 14% to 28% of the time, depending on the order of the stimuli.

The accuracy of two subjects was significantly higher for the Categorical SimChain task than for the Arbitrary SimChain task, both overall (Benedict: $\frac{P}{R} > 301, df = 2, p < .001$; Horatio: $\frac{P}{R} > 70.1, df = 2, p =< .001$) and in post-hoc pairwise comparisons ($p < .001$, corrected for multiple comparisons using the Holm-Šidák procedure). These subjects' accuracy was also significantly lower for the Concept Chain task. The accuracy of the remaining subjects was not significantly different as a function of task (Macduff: $\frac{P}{R} = 0.83, df = 2, p = .65$; Prospero: $\frac{P}{R} = 4.65, df = 2, p = .09$). According to binomial tests, Category SimChain and Concept Chain performance for all subjects (as defined by proportion of trials ending in reward) exceeded that of the bag-of-features algorithm ($p < .01$).

Figure 9 shows mean log-scaled reaction times for each list position (+/- 1 standard error). As in Experiment 1, every subject showed a significant difference in response speed as a function of list position (Benedict: $F(3, 10231) > 1373, p < .001$; Horatio: $F(3, 8391) > 345, p < .001$; Macduff: $F(3, 6649) > 408, p < .001$; Prospero: $F(3, 8298) > 1441, p < .001$). These differences displayed a substantial effect size (Benedict: $\frac{P}{R} = 0.237$; Horatio: $\frac{P}{R} = 0.334$; Macduff: $\frac{P}{R} = 0.426$; Prospero: $\frac{P}{R} = 0.333$) that dominated the proportion of variance explained. Although significant differences for task type were observed (Benedict: $F(2, 10231) > 54.4, p < .001$; Horatio: $F(2, 8391) > 1.62, p < .001$; Macduff: $F(2, 6649) > 2.24, p < .001$; Prospero: $F(2, 8298) > 32.0, p < .001$), their effect sizes were negligible (Benedict: $\frac{P}{R} = 0.006$; Horatio: $\frac{P}{R} = 0.009$; Macduff: $\frac{P}{R} = 0.001$; Prospero: $\frac{P}{R} = 0.011$). Similarly, we observed a significant interaction between these two main effects (Benedict: $F(6, 10231) > 33.8, p < .001$; Horatio: $F(6, 8391) > 21.0, p < .001$; Macduff: $F(6, 6649) > 8.78, p < .001$; Prospero: $F(6, 8298) > 9.33, p < .001$), but the effect sizes of these differences were also very small (Benedict:

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.967v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

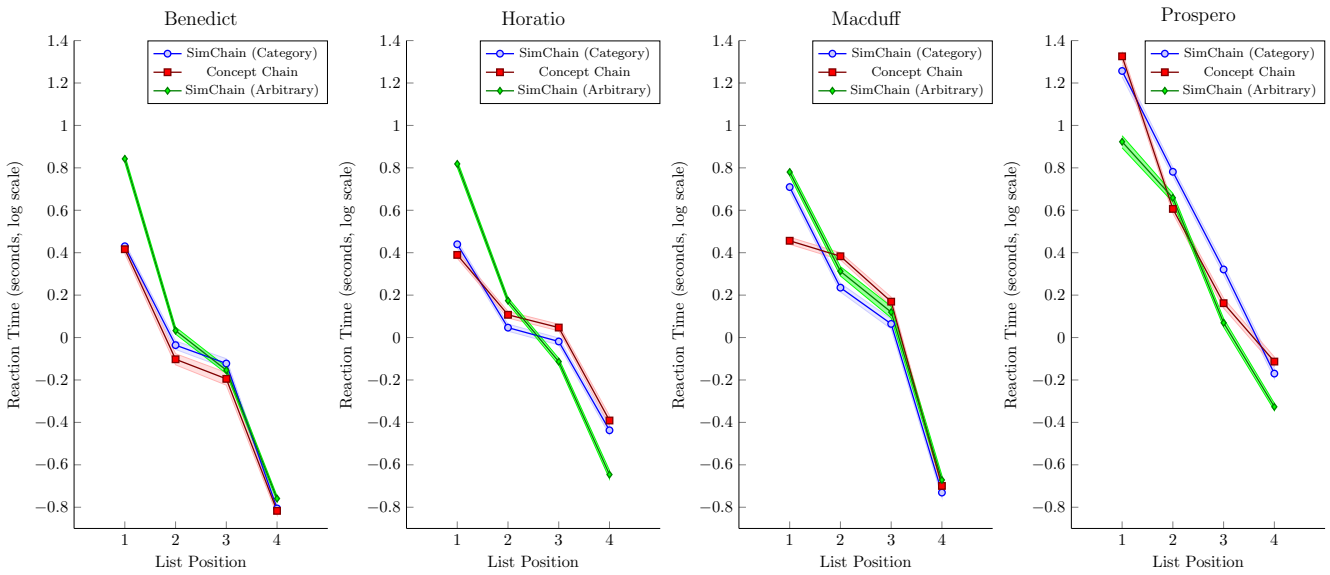

Figure 9. Average of log-scaled reaction times conditional on progress in each trial in Experiment 2. For example, List Position 1 times show to interval prior to the first touch, whereas List Position 2 times correspond to reaction times given that the first touch was correct. Plotted is response accuracy for Category SimChain trials (blue circles), Concept Chain trials (red squares), and Arbitrary SimChain trials (green diamonds), with shaded regions depicting one standard error. Included are 10237 RTs from Benedict, 8391 RTs from Horatio, 6655 RTs from MacDuff, and 8304 RTs from Prospero.

$\frac{P}{R} = 0.011$; Horatio: $\frac{P}{R} = 0.040$; Macduff: $\frac{P}{R} = 0.018$; Prospero: $\frac{P}{R} = 0.009$).

Unlike Experiment 1, performance in Experiment 2 was not uniformly better on concept lists than control lists. Although subjects performed at above chance levels in the Concept Chain task, they also performed either below or at comparable levels to their SimChain performance. During the transfer test, two monkeys displayed clear benefits of concept training. A third monkey showed mixed benefits, and the fourth treated category members as if they were arbitrary stimuli.

Although these macaques demonstrated an ability to form concepts based on highly abstract stimuli, discriminating between painterly styles was nevertheless more difficult than categorizing ecological stimuli. This may in part stem from the perceptually salient features that ecological stimuli typically possess (Marsh and MacDonald, 2008). The painting stimuli lacked discrete identifiable features on a scale that would permit such a strategy. Despite this, reaction times were very similar for each of the tasks. This fact suggests that both tasks demanded similar perceptual processing.

GENERAL DISCUSSION

Our results provide evidence that monkeys can accurately identify four simultaneously presented stimuli, belonging to ecological and stylistic groupings and selected from large and highly disparate stimulus sets. Subjects' performance cannot be explained by treating each stimulus as a percept whose categorization is subject to associative learning. This conclusion follows from a comprehensive analysis of response accuracy and reaction times, as well as from a detailed image analysis of our stimuli (provided in the online supplement) and by simulated performance using the bag-of-features classifier.

Having ruled out reinforcement learning as a sufficient explanation for performance, we propose that subjects made use of *conceptual representations* (Newen and Bartels, 2007). A conceptual representation is a non-linguistic form of conceptual learning that is more generalized and flexible than mere feature-based reinforcement learning. Three characteristics are required for a representation to qualify as conceptual:

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.967v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

identification (stimuli are consistently classified), *independence* (selection is not based on task-related discriminative cues), and *abstraction* (it is impossible to categorize stimuli by feature generalization alone). These requirements collectively provide a reasonable framework for conceptual learning that doesn't require language. Perceptual classification by way of conceptual representation therefore helps to fill the theoretical gap between reinforcement learning and linguistic concepts identified by Herrnstein (1990).

Our procedure was more challenging than any previous task used to assess concept learning because it required subjects to learn four concepts in parallel, and to then classify corresponding stimuli simultaneously. If subjects were required to identify only one concept at a time, it could be argued that some representative features of the stimulus (or of the task itself) functioned as a conditioned cue. This argument is not valid when four exemplars are presented simultaneously in a paradigm in which the subject must respond to them in a specific order before a reward is delivered. This feature, which is integral to the SimChain task, was designed to rule out associative accounts of serial learning (Terrace, 1984). The Concept Chain renders reinforcement learning even less likely because stimuli change from one trial to the next. Taken together, these variations of the SimChain task provide an effective method for addressing the problems of dichotomous training and testing described by Jensen and Altschul (2015).

The stimulus sets used in this study were larger and more diverse than those used in previous studies of concept learning in animals. Instead of making the stimuli as uniform as possible, we deliberately used diverse stimulus sets. In particular, the painting stimuli used in Experiment 2 are difficult for feature-based strategies to identify. This distinguishes them from stimuli depicting man-made objects used in earlier studies (Sigala, 2009; Fize et al., 2011), because those objects had consistent features (such as wheels or windows). The painting stimuli also satisfy the "abstraction" requirement of the conceptual representation construct.

The painting stimuli engendered lower performance (in both the monkeys and the bag-of-features algorithm) than did the ecological stimuli. The availability of easily distinguished features likely facilitates stimulus classification, which may explain higher performance in Experiment 1. It is also possible that ecologically relevant photographs have higher salience than paintings, as a result of evolutionary pressures (New et al., 2007; Fize et al., 2011; Crouzet et al., 2012). However, subjects' success in Experiment 2, as well as past studies using symbolic or artistic stimuli (Schrier et al., 1984; Matsukawa et al., 2001; Watanabe, 2013), strongly suggest that discrete features such as eyes are not a necessary prerequisite for concept formation.

Other studies have found that macaques reliably and rapidly classify images that have been systematically degraded (Macé et al., 2010; Basile and Hampton, 2013). This undermines speculations that classification is exclusively feature-driven and reliant on visual search. Reaction times in our study were consistent with this literature: Trial-to-trial changes in the stimuli did not yield qualitatively different response speeds than did the SimChain task. Reaction times were also similar when comparing ecological and painterly stimuli. The consistency of reaction times suggests a common cognitive architecture for both perceptual and conceptual learning (Goldstone and Barsalou, 1998), which challenges the assumptions underlying many skeptical claims about concept formation in animals.

A better account of this consistency is offered by *grounded cognition* (Barsalou, 2008), wherein an animal's various conceptual aptitudes are grounded in and arise from common perceptual machinery. Recent animal studies of the neural architecture of conceptual representations reveal highly specialized and modality-specific networks. For example, the location and organization of networks dedicated to visual representations are pre-specified (Srihasam et al., 2014). Furthermore such networks stretch back into the sensory-motor regions of the modality through which the concept was initially learned (Martin, 2008). Although parsimony is often invoked when arguing in favor of reinforcement learning, the neural networks responsible for stimulus classification must also be taken into account. In an evolutionary context, skeptics

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.967v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

must acknowledge evidence that aptitudes for these forms of learning are built on the foundations of the brain's perceptual processes. Given similarities in performance, it seems unlikely that Experiments 1 and 2 recruited distinct neural networks. More likely, performance in both depended on the same conceptual machinery.

Given the positive results reported in the literature's most rigorous studies of animal concept learning (e.g. Bhatt et al., 1988), it should come as no surprise that non-human primates can flexibly classify stimuli according to abstract properties. Much as an average museum visitor does not possess technical language to describe a painting, but may still have a hunch who painted it, monkeys need not understand the subject matter, nor even the art of painting, in order to be able to distinguish the work of one painter from another. Our study highlights the need for additional research on how non-verbal subjects learn concepts that are not linguistic, as conceptual representations of this sort are likely held by humans (Younger, 2010; Fize et al., 2011). Nevertheless, a gap remains in accounting for the flexibility of animals' perceptual concepts, following from the undefined distinction between the most advanced open-ended categorization and simplest abstract concepts (Roberts, 1996).

Given the positive results we obtained with 4-item Concept Chains, we anticipate that future investigations will benefit from more demanding test procedures. Using larger stimulus sets will help rule out memorization. Training subjects on a wider range of categories (e.g. by using 7-item SimChains Terrace et al., 2003) would further reduce problems inherent in binary testing. Using more difficult tasks, as well as demonstrating transfer of knowledge between tasks (e.g. Jensen et al., 2013a) should also clarify the nature of subjects? representations, including their limitations. Such studies should help to better define concepts, and to distinguish abstract relations from more basic discriminative processes (Zentall et al., 2002).

ACKNOWLEDGMENTS

The authors wish to thank Adebayo Adesomo, Farhana Begum, Shangshang Chen, Daniel Deihle, and Rachelle Meyer for their assistance with the monkeys, and Vincent Ferrera for feedback on an earlier version of this manuscript. This work was supported by US National Institute of Health, grant 5R01MH081153-06 awarded to H.T.

AUTHOR CONTRIBUTIONS

D.A., G.J., and H.S. conceived the experiments. D.A. and G.J. wrote the task software, acquired data, and performed analyses. D.A., G.J., and H.S. wrote the paper.

REFERENCES

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, 59:617–645.

Basile, B. M. and Hampton, R. R. (2013). Monkeys show recognition without priming in a classification task. *Behavioural Processes*, 93:50–61.

Bhatt, R. S., Wasserman, E. A., Reynolds, W. F., and Knauss, K. S. (1988). Conceptual behavior in pigeons: Categorization of both familiar and novel examples from four classes of natural and artificial stimuli. *Journal of Experimental Psychology: Animal Behavior Processes*, 14:219–234.

Chater, N. and Heyes, C. (1994). Animal concepts: Content and discontent. *Mind and Language*, 9:209–246.

Crouzet, S. M., Joubert, O. R., Thorpe, S. J., and Fabre-Thorpe, M. (2012). Animal detection precedes access to scene category. *PLOS ONE*, 7:e51471.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.967v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

Escher, M. C. (1989). *Escher on Escher: Exploring the Infinite*. Harry N. Adams, New York, NY.
Fize, D., Cauchoix, M., and Fabre-Thorpe, M. (2011). Humans and monkeys share visual representations. *Proceedings of the National Academy of Sciences of the USA*, 108:7635–7640.
Flemming, T. M., Thompson, R. K., and Fagot, J. (2013). Baboons, like humans, solve analogy by categorical abstraction of relations. *Animal Cognition*, 16:519–524.
Goldstone, R. L. and Barsalou, L. W. (1998). Reuniting perception and conception. *Cognition*, 65:231–262.
Griffin, G., Holub, A. D., and Perona, P. (2007). The Caltech 256. Technical Report CNS-TR-2007-001, California Institute of Technology.
Herrnstein, R. J. (1990). Level of stimulus control: A functional approach. *Cognition*, 37:133–166.
Herrnstein, R. J., Loveland, D. H., and Cable, C. (1976). Natural concepts in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, 2:285–302.
Huber, L. (2000). Generic perception: Open-ended categorization of natural classes. In Fagot, J., editor, *Picture Perception in Animals*, pages 219–261. Psychology Press, Hove, East Sussex.
Jensen, G. (2013). Closed-form estimation of multiple change-point models. *PeerJ Preprints*, 1:e90v3.
Jensen, G. and Altschul, D. (2015). Two perils of binary categorization: Why the study of concepts can't afford true/false testing. *Frontiers in Psychology*, 6:Article 168.
Jensen, G., Altschul, D., Danly, E., and Terrace, H. S. (2013a). Transfer of a serial representation between two distinct tasks by rhesus macaques. *PLOS ONE*, 8:e70285.
Jensen, G., Ward, R. D., and Balsam, P. D. (2013b). Information: Theory, brain, and behavior. *Journal of the Experimental Analysis of Behavior*, 100:408–431.
Jitsumori, M. and Delius, J. D. (2001). Object recognition and object categorization in animals. In Matsuzawa, T., editor, *Primate Origins of Human Cognition and Behavior*, pages 269–293. Springer, New York, NY.
Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer, New York, NY.
Katz, J. S., Wright, A. A., and Bodily, K. D. (2007). Issues in the comparative cognition of abstract-concept learning. *Comparative Cognition and Behavior Reviews*, 2:79–92.
Lea, S. E. G. (1984). In what sense do pigeons learn concepts? In Roitblat, H. L., Bever, T. G., and Terrace, H. S., editors, *Animal Cognition*, pages 263–276. Erlbaum, Hillsdale, NJ.
Lee, S. M., Xin, J. H., and Westland, S. (2005). Evaluation of image similarity by histogram intersection. *Color Research and Application*, 30:265–274.
Macé, M. J.-M., Delorme, A., Richard, G., and Fabre-Thorpe, M. (2010). Spotting animals in natural scenes: efficiency of humans and monkeys at very low contrasts. *Animal Cognition*, 13:405–418.
Marsh, H. L. and MacDonald, S. E. (2008). The use of perceptual features in categorization by orangutans (*Pongo abelli*). *Animal Cognition*, 11:569–585.
Martin, A. (2008). The representation of object concepts in the brain. *Annual Review of Psychology*, 58:25–45.
Matsukawa, A., Inoue, S., and Jitsumori, M. (2001). Pigeon's recognition of cartoons: effects of fragmentation, scrambling, and deletion of elements. *Behavioural Processes*, 65:25–34.
Miller, E. K., Nieder, A., Freedman, D. J., and Wallis, J. D. (2003). Neural correlates of categories and concepts. *Current Opinion in Neurobiology*, 13:198–203.
New, J., Cosmides, L., and Tooby, J. (2007). Category-specific attention for animals reflects ancestral priorities, not expertise. *Proceedings of the National Academy of Sciences of the USA*, 104:16598–16603.
Newen, A. and Bartels, A. (2007). Animal minds and the possession of concepts. *Philosophical Psychology*, 20:283–308.
Nilsback, M.-E. and Zisserman, A. (2008). Automated flower classification over a large number of classes.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.967v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*.
O'Hara, S. and Draper, B. A. (2011). Introduction to the bag of features paradigm for image classification and retrieval. *arXiv Preprints*, page 1101.3354.
Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. V. (2012). Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*.
Roberts, W. A. (1996). Stimulus generalization and hierarchical structure in categorization by animals. *Advances in Psychology*, 117:35–54.
Roberts, W. A. and Mazmanian, D. S. (1996). Concept learning at different levels of abstraction by pigeons, monkeys, and people. *Journal of Experimental Psychology: Animal Behavior Processes*, 14:247–260.
Schrier, A. M., Angarella, R., and Povar, M. L. (1984). Studies of concept formation by stumptailed monkeys: Concepts humans, monkeys, and letter *A*. *Journal of Experimental Psychology: Animal Behavior Processes*, 10:564–584.
Schrier, A. M. and Brady, P. M. (1987). Categorization of natural stimuli by monkeys (*Macaca mulatta*): Effects of stimulus set size and modification of exemplars. *Journal of Experimental Psychology: Animal Behavior Processes*, 13:136–143.
Sigala, N. (2009). Natural images: A lingua franca for primates? *Open Neuroscience Journal*, 3:48–51.
Srihasam, K., Vincent, J. L., and Livingstone, M. S. (2014). Novel domain formation reveals proto-architecture in inferotemporal cortex. *Nature Neuroscience*, 17:1776–1783.
Terrace, H. S. (1984). Simultaneous chaining: The problem it poses for traditional chaining theory. In Commons, M. L., Herrnstein, R. J., and Wagner, A. R., editors, *Quantitative Analyses of Behavior: Discrimination Processes*, pages 115–138. Ballinger, Cambridge, MA.
Terrace, H. S., Son, L. K., and Brannon, E. M. (2003). Serial expertise of rhesus macaques. *Psychological Science*, 14:66–73.
Thurstone, L. L. (1919). The learning curve equation. *Psychological Monographs: General and Applied*, 26:1–51.
Vogels, R. (1999). Categorization of complex visual images by rhesus monkeys. part 2: single-cell study. *European Journal of Neuroscience*, 11:1239–1255.
Vonk, J. (2013). Matching based on biological categories in orangutans (*Pongo abelii*) and a gorilla (*Gorilla gorilla gorilla*). *PeerJ*, 1:e158.
Vonk, J. and MacDonald, S. E. (2002). Natural concepts in a juvenile gorilla (*Gorilla gorilla gorilla*) at three levels of abstraction. *Journal of the Experimental Analysis of Behavior*, 78:315–332.
Wasserman, E. A., Fagot, J., and Young, M. E. (2001). Same-different conceptualization by baboons (*Papio papio*): The role of entropy. *Journal of Comparative Psychology*, 115:42–52.
Watanabe, S. (2013). Preference for and discrimination of paintings by mice. *PLOS ONE*, 8:e65335.
Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., and Perona, P. (2010). Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology.
Young, M. E. and Wasserman, E. A. (2002). The pigeon's discrimination of visual entropy: A logarithmic function. *Animal Learning and Behavior*, 30:306–314.
Younger, B. A. (2010). Categorization and concept formation in human infants. In Mareschal, D., Quinn, P. C., and Lea, S., editors, *The Making of Human Concepts*, pages 245–263. Oxford University Press, New York, NY.
Zentall, T. R., Galizio, M., and Critchfield, T. S. (2002). Categorization, concept learning, and behavior analysis: An introduction. *Journal of the Experimental Analysis of Behavior*, 78:237–248.
Zentall, T. R., Wasserman, E. A., Lazareva, O. F., and Rattermann, M. J. (2008). Concept learning in animals. *Comparative Cognition and Behavior Reviews*, 3:13–45.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.967v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

APPENDIX: STIMULUS ANALYSIS

Although comparative studies of primate cognition very often use photographic stimuli, systematic analyses of the stimuli are rarely undertaken. This is unfortunate, because skepticism about surprising results often relies on speculation about stimulus characteristics that *might* have been used as discriminative cues. A subject might, for example, be suspected of identifying pictures of birds solely on the basis of a blue backdrop at the top of the images (i.e. the sky). Rather than speculate about the putative properties of the stimuli, a comprehensive analysis can determine whether bird stimuli contain disproportionate amounts of blue.

Our analysis of stimulus images focuses on low-level features like image entropy and color histograms. This approach has the advantage of being entirely automatic and replicable, which is especially important given the large number of stimuli we employed. The overarching question that these analyses seek to inform is this: "To what extent are low-level properties sufficient to categorize stimuli correctly?" The more diverse the stimuli in each grouping, the more difficult it is to specify criteria for category inclusion. At the same time, the more each of the categories resembles the others overall, the more difficult it is to specify criteria for category exclusion.

Stimuli

The stimuli used in this study were selected in a fashion that differs from the conventions used in typical psychophysical experiments. Rather than select stimuli according to strict inclusion criteria, or modifying images before use (e.g., turning them grayscale or giving them uniform spectra), we included images solely on the basis of the question, "Is this a picture of X?" For example, our photographs of people included both extreme close-ups of faces and wide-angle views of crowds. We also included both color and black-and-white images.

Figures 10 depict a representative set of exemplars for each of the conceptual categories used in Experiment 2. For reasons associated with image and likeness rights, exemplars from Experiment 1 are not included. However, representative images may be obtained from the Caltech-UCSD Birds 200 Dataset (Welinder et al., 2010) for the 'birds' category, from the Oxford-IIIT Pet Dataset (Parkhi et al., 2012) for the 'cats' category, from the Oxford 102 Category Flower Dataset (Nilsback and Zisserman, 2008) for the 'flowers' category, and from the Caltech 256 Dataset (Griffin et al., 2007) for the 'people' category. In all cases, these stimuli were selected for this study. Subjects with prior experience using the SimChain paradigm were not previously exposed to these specific images.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.967v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

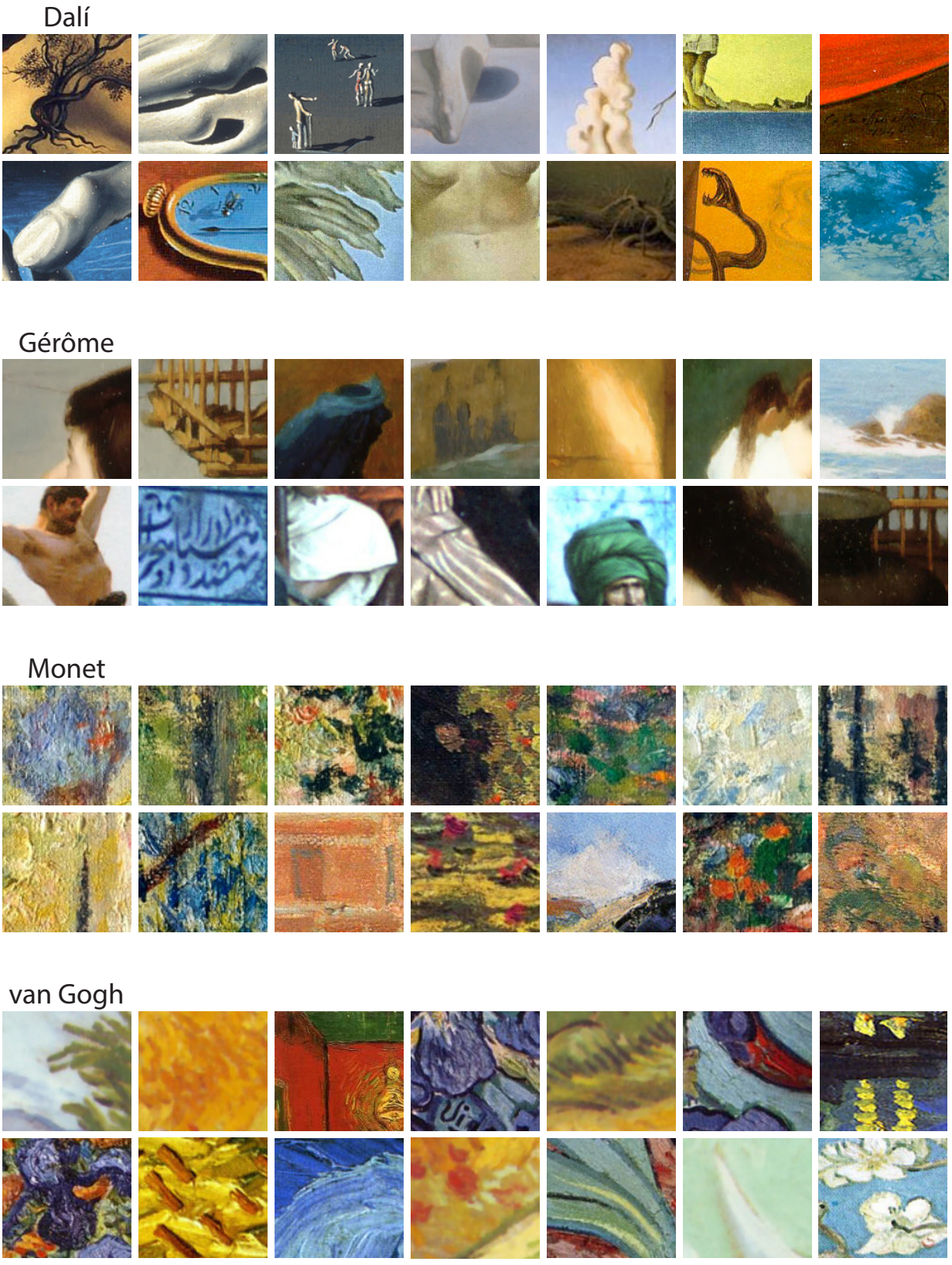

Figure 10. Exemplars of the stimuli drawn from the works of four painters, used in Experiment 2.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.967v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

Stimulus Analysis: Pixel Entropy

Prior research has shown that primates possess the ability to discriminate stimuli based on visual entropy Flemming et al. (2013); Wasserman et al. (2001), an ability also demonstrated in pigeons Young and Wasserman (2002). Because the entropy estimation can be done mechanically by simple systems, doing so falls considerably short of the criteria for a "conceptual representation." Consequently, an analysis of pixel entropies gives an idea of whether the sets of stimuli differ sufficiently to be discriminated on that basis.

Here, pixel entropy is taken to be the Shannon entropy Jensen et al. (2013b), computed over all possible combinations of red, blue, and green intensities:

$$H = \sum_{r=0}^{255} \sum_{g=0}^{255} \sum_{b=0}^{255} p\left(r,g,b\right) \log_2 \left(p\left(r,g,b\right)\right) \tag{5}$$

The maximum possible entropy H that a bitmap image could possibly display is 24, provided each of the $256 \times 256 \times 256$ pixel values appears equally. However, such an entropy would require a 4096×4096 pixel image, much larger than our stimuli. Because our stimuli were only 140×130 pixels in size, the highest possible entropy that a color stimulus could possess was 14.15 bits. Grayscale images had a maximum entropy of 8 bits.

Figures 11 and 12 show kernel density estimates of the distributions of pixel entropies displayed in Experiments 1 and 2, respectively, as well as each distribution's quartiles. In general, stimuli tended to show high entropies of between 12 and 14 bits, such that a 13-bit image could easily belong to any of the categories. However, the stimuli used in Experiment 1 do show clear distributional differences. For example, many more of the images of birds have entropies below twelve than the other stimuli, while the images of flowers routinely have higher entropies than the other stimuli.

The stimuli used in Experiment 2 tend to have higher entropies overall than those in Experiment 1. Here, too, however, there are notable similarities. Dalí and Gérôme both resemble one another closely, as do Monet and van Gogh, but these two clusters appear distinct from one another. Importantly, however, because stimuli in each of these pairings are distributed so similarly, it would be very difficult for subjects to distinguish each group precisely on the basis of pixel entropy alone.

We do not rule out the possibility that pixel entropy facilitated identification in some fashion. This analysis is merely intended to demonstrate that pixel entropy alone would not have been sufficient to precisely classify each stimulus.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.967v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

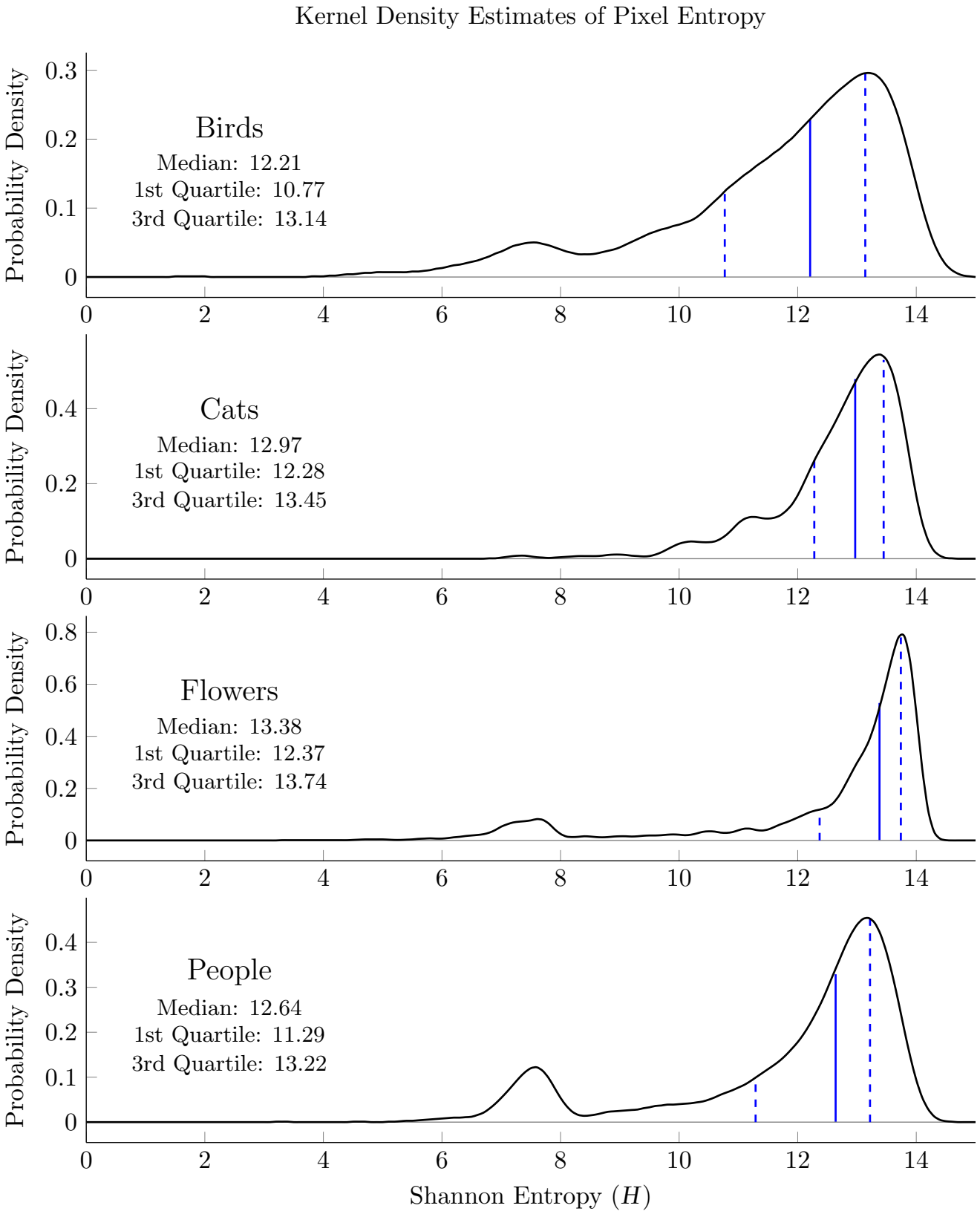

Figure 11. Kernel density estimates of pixel entropy in the four categories used for Experiment 1. The median image is indicated by the solid blue line, while the first and third quartiles are indicated by the blue dashed lines.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.967v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

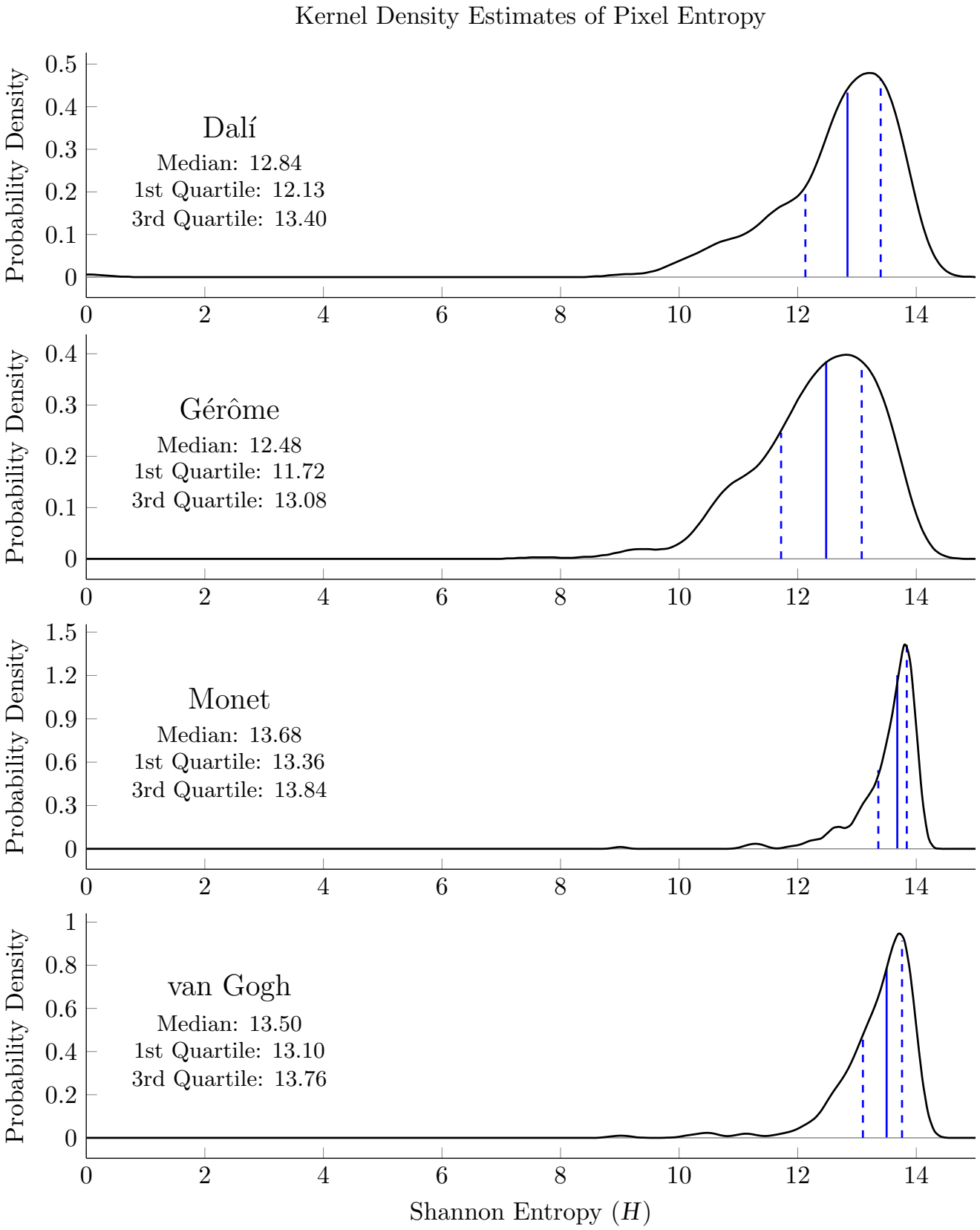

Figure 12. Kernel density estimates of pixel entropy in the four categories used for Experiment 1. The median image is indicated by the solid blue line, while the first and third quartiles are indicated by the blue dashed lines.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.967v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

Stimulus Analysis: HSV Histograms

Another method by which images can be compared is on their HSV distributions. Just as each image may be represented as a collection of pixels that have values of red, blue, and green, each pixel may also be represented by the orthogonal dimensions of hue, saturation, and value (the last corresponding to the luminosity of the pixel). HSV histograms are often more subjectively informative than RGB histograms, as they are better at revealing effects such as tint, brightness, and color intensity Lee et al. (2005).

For this analysis, the histograms of hue, saturation, and value were obtained for each stimulus. Then, these stimuli were sorted according to similarity by performing a principal component analysis Jolliffe (2002) and ordering the histograms according to the first component. This yields a 3D map of frequencies across stimuli, in which each row represents a single stimulus and each column represents a particular index in the histogram.

Figure 13 plots this multi-image histogram as a heat map for the hues of all stimuli in Experiment 1. In addition to the histograms for each individual image, Figure 13 also plots the marginal frequencies across *all* stimuli in each category. Here, we can see quite clearly that the different categories reliably have properties that can be used to distinguish one category from the next. Pictures of birds very frequently have green and cyan elements (because of leaves or sky), and flowers have a greater representation of yellow and purple. Photographs of people tend to be more reddish, while cats tend to be more orange.

Note that the apparently ”blank bands” visible in these heat maps are black-and-white images. Since a black-and-white image cannot reasonable be described as having a particular hue, the frequency distribution for those images were uniform.

Figure 14 plots the histogram for saturation of stimuli in Experiment 1, and here, too, patterns of differ visibly. Photographs of flowers are typically highly saturated, while photos of cats and birds tend to have low saturation. However, an examination of the distributions of individual stimuli suggest that there is an overall level of heterogeny in most cases, as evidenced by the lack of consistent vertical bands in the heat maps.

Figure 15 plots the histogram for value (i.e. brightness) of stimuli in Experiment 1, showing clear differences. Flowers and people tend to be spread across the range, while birds and cats tend to cluster toward the center. As in the case of saturation, there is a great deal of variation across stimuli, such that these would not be strongly selective signals.

Figure 16 plots the histogram for hue of the painting stimuli in Experiment 2, and a much greater degree of uniformity is observed here than in Experiment 1. In all cases, painters favored colors in the orange-yellow and cyan ranges. This similarity across painters, combined with the heterogeny of the images (in which many had no blue to them at all) ensures that hue could not be used as a reliable cue in Experiment 2.

Figure 17 plots the histogram for saturation of the painting stimuli in Experiment 2. Here, heterogeny dominated, with all painters having images that were spread across the full range of saturations. While there were some differences (Dalí tended to be the most likely to have highly saturated colors, for example), the spread across the range prevented saturation from being a reliable cue.

Figure 18 plots the histogram for value of the painting stimuli in Experiment 2. As in the case of saturation, the painters were highly heterogeneous, tending to favor intermediary values. This is unsurprising, as artists routinely avoid using pure white and pure black, instead favoring intermediate values that give an impression of contrast Escher (1989).

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.967v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

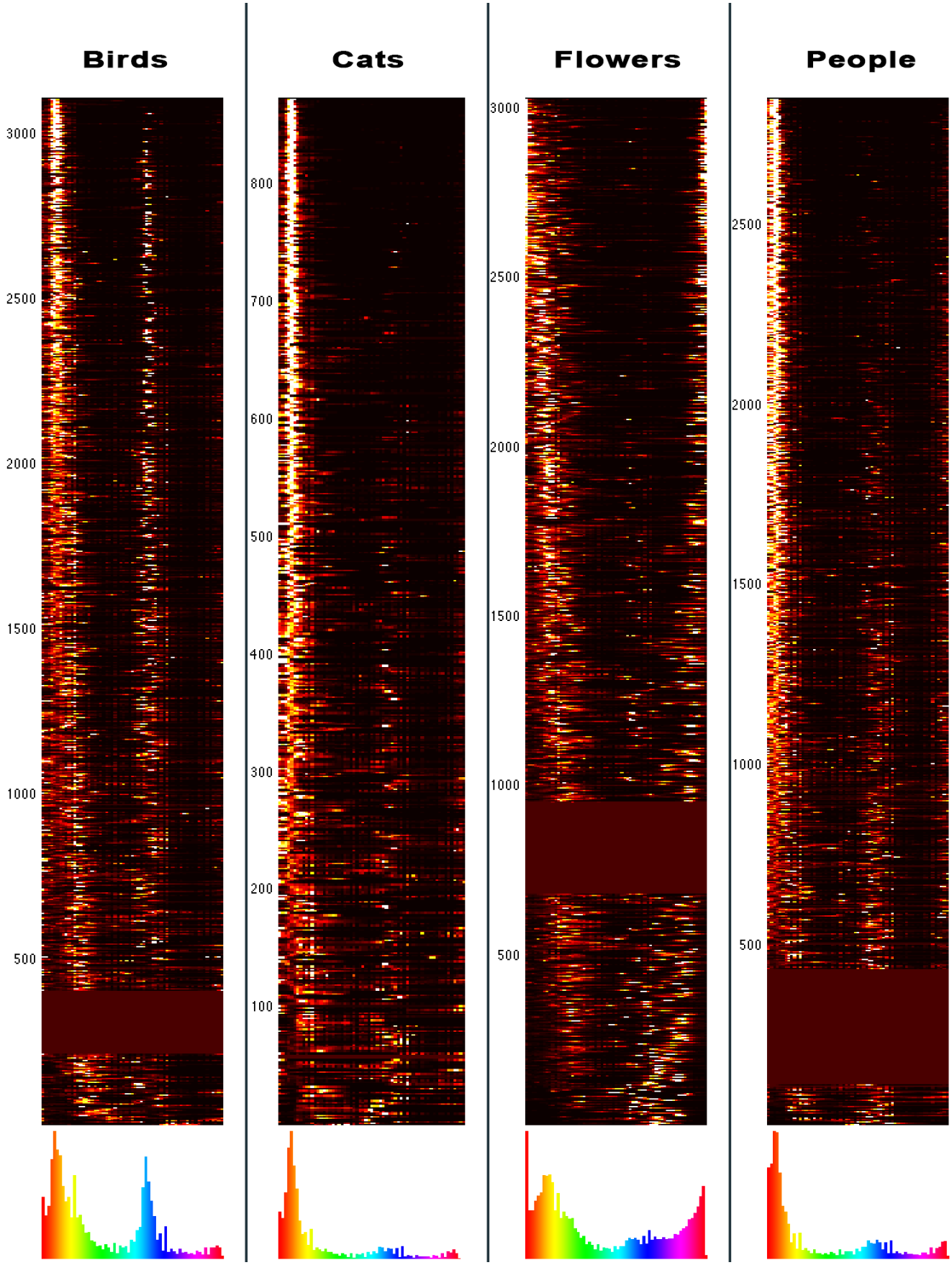

Figure 13. Hue histograms for each stimulus used in Experiment 1, sorted using principal component analysis. The histogram at the bottom depicts the marginal frequency across all images.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.967v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

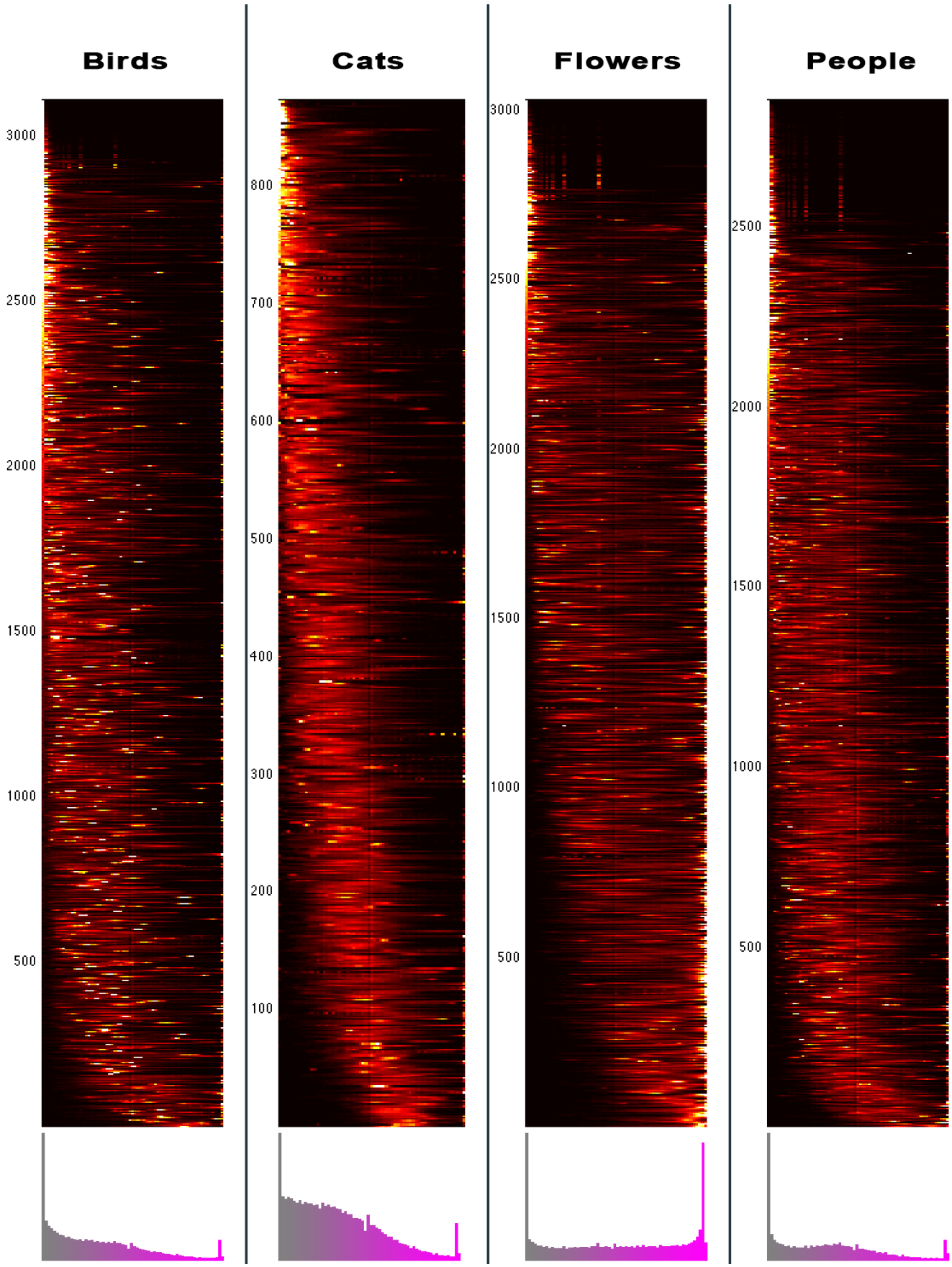

Figure 14. Saturation histograms for each stimulus used in Experiment 1, sorted using principal component analysis. The histogram at the bottom depicts the marginal frequency across all images.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.967v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

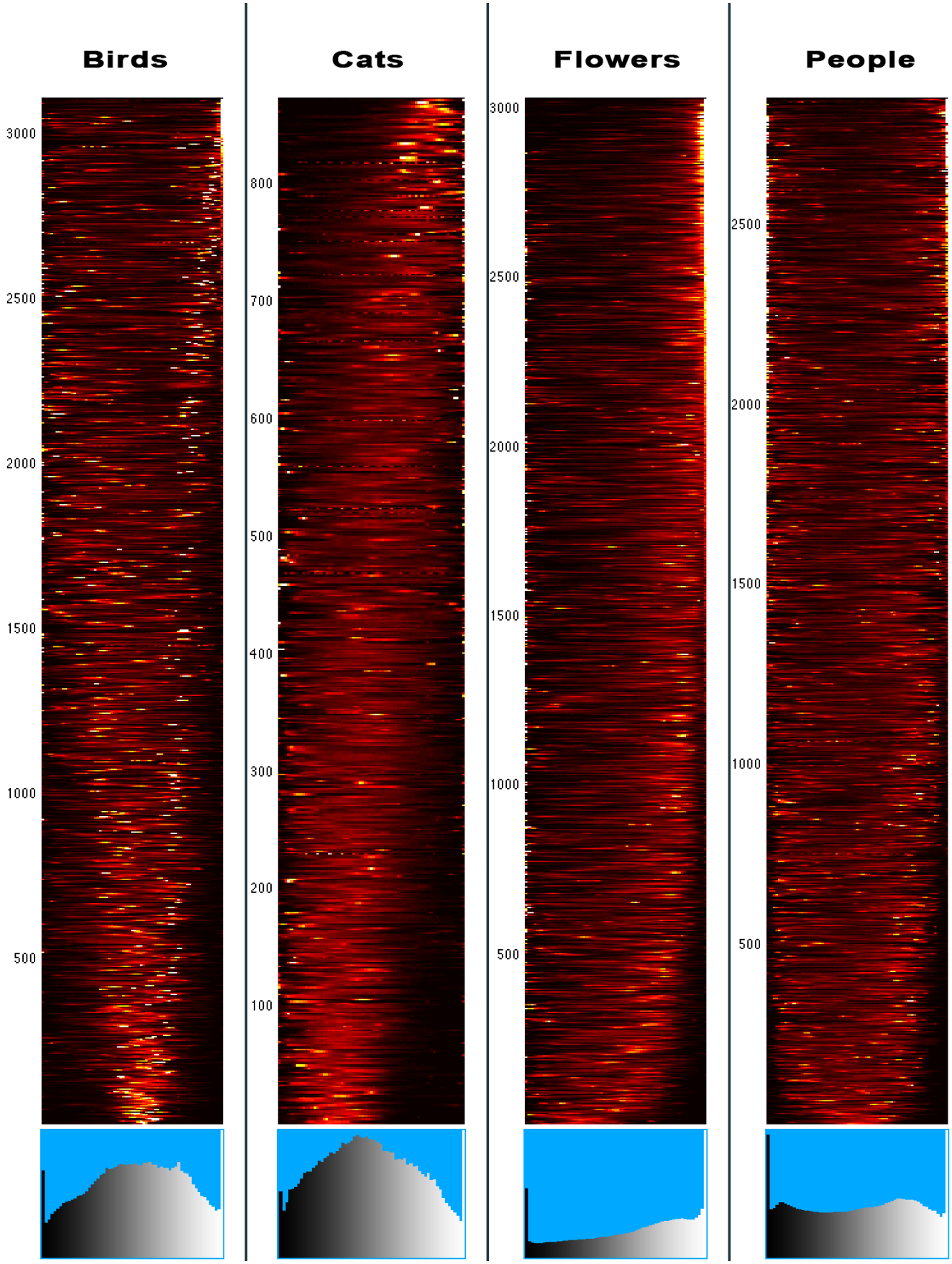

Figure 15. Value histograms for each stimulus used in Experiment 1, sorted using principal component analysis. The histogram at the bottom depicts the marginal frequency across all images.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.967v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

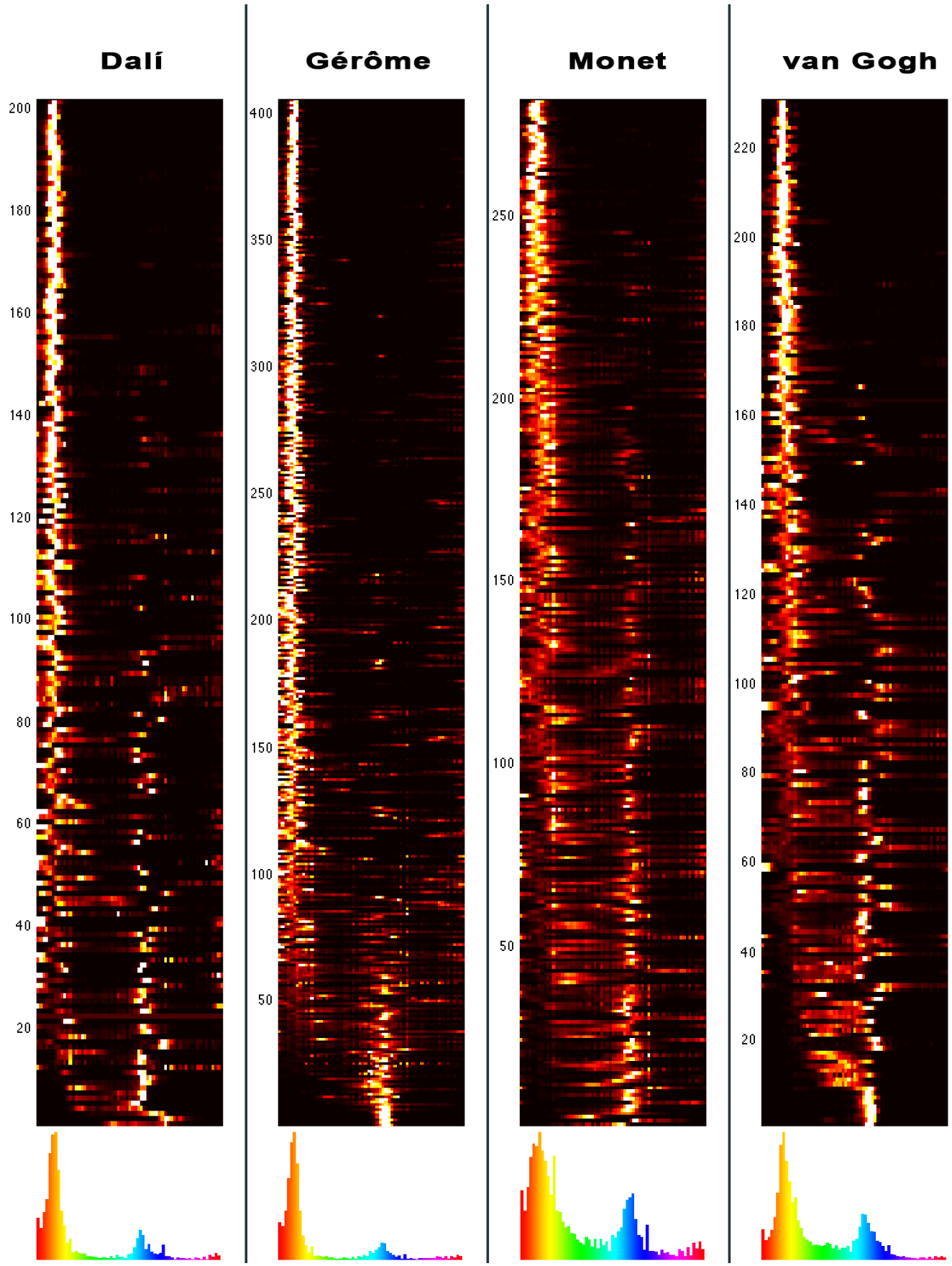

Figure 16. Hue histograms for each stimulus used in Experiment 2, sorted using principal component analysis. The histogram at the bottom depicts the marginal frequency across all images.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.967v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

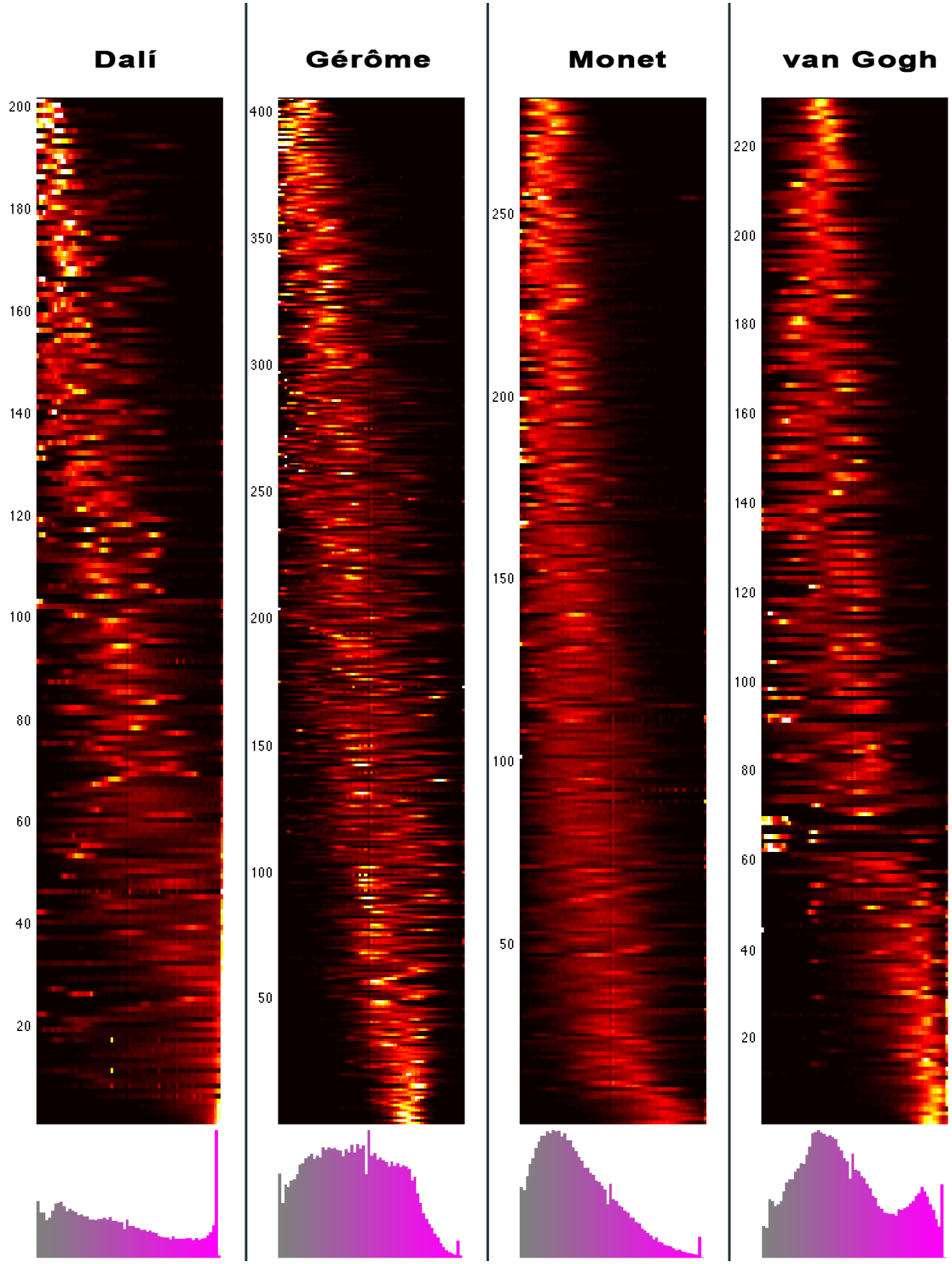

Figure 17. Saturation histograms for each stimulus used in Experiment 2, sorted using principal component analysis. The histogram at the bottom depicts the marginal frequency across all images.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.967v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

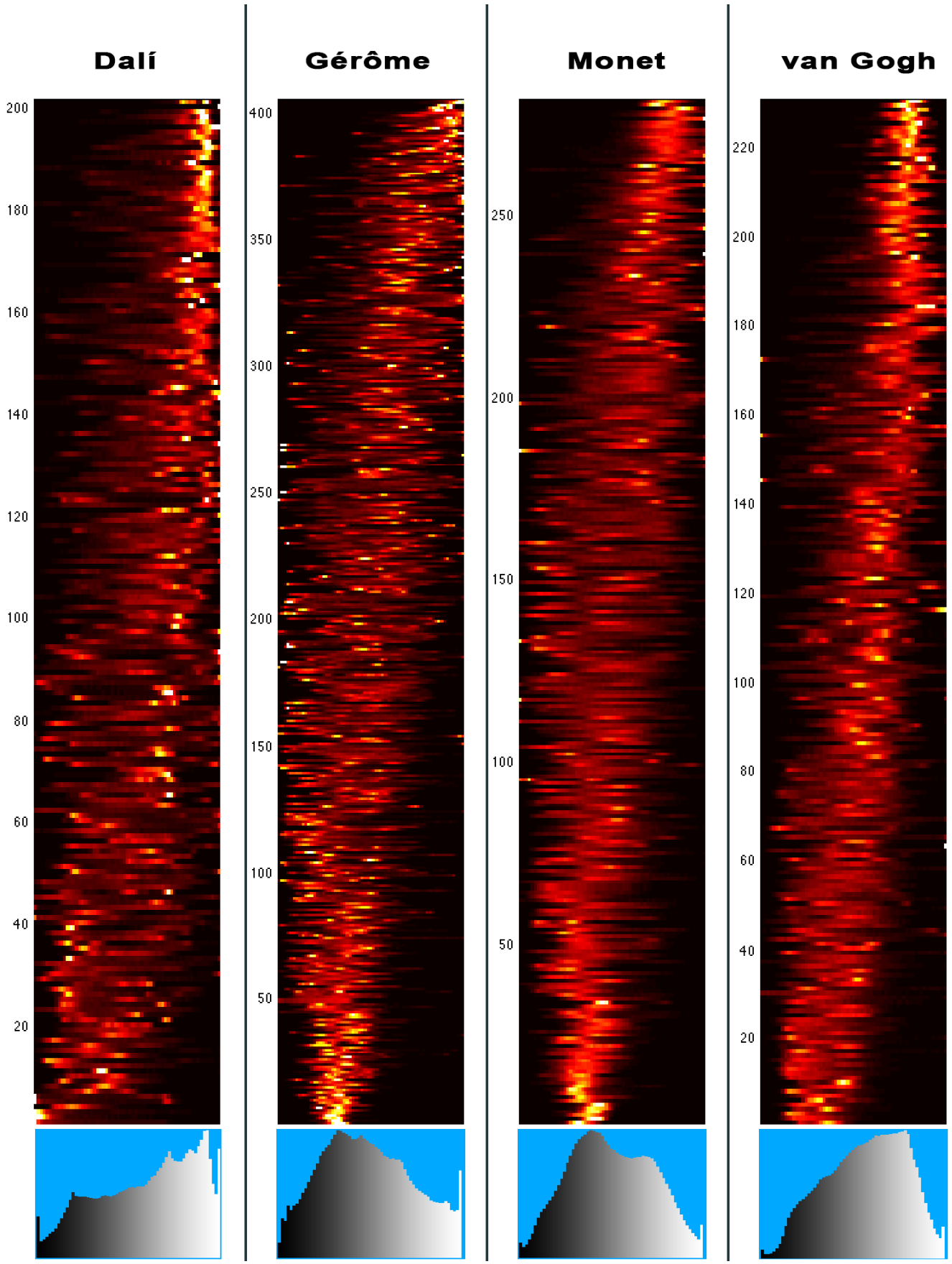

Figure 18. Value histograms for each stimulus used in Experiment 2, sorted using principal component analysis. The histogram at the bottom depicts the marginal frequency across all images.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.967v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015