TOWARDS RIGOROUS STUDY OF ARTISTIC STYLE:
A NEW PSYCHOPHYSICAL PARADIGM

Holly E. Gerhard\textsuperscript{1,2} and Matthias Bethge\textsuperscript{1-3}

\textsuperscript{1}Werner Reichardt Centre for Integrative Neuroscience
University of Tübingen, Tübingen, Germany

\textsuperscript{2}Computational Vision and Neuroscience Group
Max Planck Institute for Biological Cybernetics, Tübingen, Germany

\textsuperscript{3}Institute of Theoretical Physics
University of Tübingen, Tübingen, Germany

In press at Art & Perception

Keywords: artistic style, visual expertise, image statistics, drawings, Pieter Bruegel the Elder, stylometrics, texture discrimination

Contact:
Holly E. Gerhard
Centre for Integrative Neuroscience
AG Bethge
Otfried-Müller-Str. 25
72076 Tübingen, Germany

Email: holly.gerhard@bethgelab.org
ABSTRACT

What makes one artist’s style so different from another’s? How do we perceive these differences? Studying the perception of artistic style has proven difficult. Observers typically view several artworks and must group them or rate similarities between pairs. Responses are often driven by semantic variables, such as scene type or the presence/absence of particular subject matter, which leaves little room for studying how viewers distinguish a Degas ballerina from a Toulouse-Lautrec ballerina, for example. In the current paper, we introduce a new psychophysical paradigm for studying artistic style that focuses on visual qualities and avoids semantic categorization issues by presenting only very local views of a piece, thereby precluding object recognition. The task recasts stylistic judgment in a psychophysical texture discrimination framework, where visual judgments can be rigorously measured for trained and untrained observers alike. Stimuli were a dataset of drawings by Pieter Bruegel the Elder and his imitators studied by the computer science community, which showed that statistical analyses of the drawings’ local content can distinguish an authentic Bruegel from an imitation. Our non-expert observers also successfully discriminated the authentic and inauthentic drawings and furthermore discriminated stylistic variations within the categories, demonstrating the new paradigm’s feasibility for studying artistic style perception. At the same time, however, we discovered several issues in the Bruegel dataset that bear on conclusions drawn by the computer vision studies of artistic style.

[228 words]
1. INTRODUCTION

A visual artist’s style is defined by several choices made during the creative process, relating to composition, color palette, subject matter, and textural qualities to name just a few. These choices determine our interpretations and responses to a piece in ways that are often difficult to describe, even for those trained in art criticism. An equally difficult challenge is how to study art perception with the rigor of the scientific method, which seems categorically at odds with the freedom of artistic expression and interpretation. Nonetheless, several vision scientists have made admirable strides in measuring the aesthetic experience using rigorous scientific methods. Augustin and Wagemans (2012) provide a recent overview. In this paper, we introduce a new paradigm for studying the perceptual impact of finer scale details in artworks, i.e. those aspects of style related to texture or local image statistics. In Fig. 1 we contrast the traditional way of viewing an artwork globally with a set of local zoomed-in views that focus on the piece’s textural content.

----- Fig. 1 about here -----

In general, studying how people perceive the purely visual aspects of artistic style poses several challenges. For example, in an experiment where observers rated how similar various pairs of paintings were, a multi-dimensional scaling analysis revealed that two of the dimensions along which observers judged the paintings were the inclusion of human beings in the foreground and the presence of a body of water (Graham et al., 2010). In another study where viewers grouped paintings into clusters based on their style, the authors failed to find a set of purely visual image descriptors that correlated well with human classifications, suggesting a large contribution of semantic judgments beyond the power of current image descriptors (Wallraven et al., 2008). Similarly, in a study of how experts versus untrained observers group modern paintings, only experts categorized pieces according to line style and painting technique (Augustin & Leder, 2006), suggesting that non-experts do not judge these aspects of style although they certainly might also perceive such differences.

We therefore propose that local views of artworks such as Fig. 1B be used as stimuli if one wants to avoid the effects of recognizable subject matter and study the purely perceptual impact of finer textural details, even with non-expert observers. Depending on a particular study’s goal, the experimenter can combine such a local texture stimulus with any of the various tasks used to study art perception, such as aesthetic rating, similarity rating, or grouping.
In our experiments we chose to use a forced choice texture discrimination task, where separate textures were pitted against each other for comparison. Each texture was comprised of image patches from one artwork only and therefore represented a single artwork’s textural qualities. On a trial, observers viewed three textures like those in Fig. 2: a reference in the middle flanked by two comparisons. The task was to identify which of the two flankers was from the same category as the reference. We chose to use such a forced choice task—in which observers simultaneously view and compare multiple artworks to form a judgment—rather than asking observers to judge each individual artwork separately as it is well established that perceptual judgments of a single stimulus in isolation are conflated with individual differences in response criteria, biases, and trial ordering effects. Forced choice judgments, on the other hand, provide a purer measure of perceptual sensitivity separate from decision making processes. (See Green & Swets, 1966 for a technical explanation of why forced choice discrimination tasks are bias-free compared to “yes-no” tasks, i.e. those where stimuli are judged individually.) In our task, observers must compare the middle texture to each flanker and make a binary decision, i.e. a “forced choice,” about which flanker comes from the same category.

Unbeknownst to the observers, the two categories were authentic or inauthentic drawings by Pieter Bruegel the Elder (Table 1) from a dataset used in previous statistical analyses of artistic style (Lyu et al., 2004; Rockmore et al., 2006; Hughes et al., 2010 & 2012). We specifically chose not to tell observers that the stimuli were made from artworks in order to measure their purely perceptual judgments without any influence of their personal definitions or idiosyncratic expectations about artistic style.

Art historians have dramatically refined and narrowed Bruegel’s oeuvre of drawings in recent years due to a combination of forensic and stylistic analyses (Mielke, 1996; Royalton-Kisch, 1998; Metropolitan Museum, 2001). The drawings studied here are thirteen landscapes from a Pieter Bruegel the Elder exhibition catalogued by the Metropolitan Museum of Art in New York (2001) that included authentic pieces along with imitations removed from Bruegel’s oeuvre and copies made by his followers.

We chose to study this dataset because a recent analysis showed that the drawings can be discriminated using a sparse coding model (Hughes et al, 2010), a kind of probabilistic image model that has been compared with early visual processing stages in primate cortex (Olshausen & Field, 1996). Our goal was to measure human discrimination performance with
the same dataset and the same kind of local image samples analyzed by the algorithm. We had a strong prediction that even observers untrained in art criticism, i.e. non-experts, would be able to perform this task successfully. We have previously shown that human observers are far more sensitive to local natural image variations than the independent components analysis model, which is similar to sparse coding (Gerhard et al., 2013), demonstrating the greater power of the human visual system at detecting statistical image features.

The focus of this paper is to introduce a novel paradigm for studying artistic style perception psychophysically, which could reveal new insights on aesthetic experience by focusing on local image content.

2. RESULTS

2.1. Main Experiment

In Experiment 1, seven observers naïve to the purpose of the experiment and lacking artistic training performed the texture discrimination experiment. They performed the task illustrated in Fig. 2, where authentic and inauthentic Pieter Bruegel the Elder drawings (Table 1) were pitted against each other. The drawings are the subject of several statistical analyses of artistic style (Lyu et al., 2004, Hughes et al., 2010 & 2012; Rockmore et al., 2006), and we designed our task for comparison to the sparse coding discriminator in particular (Hughes et al., 2010). The image patches comprising the textures corresponded to 32×32 pixel regions in the original scans of the drawings (approximately equal to 31×31 mm square patches in the drawings), which was the largest patch size studied in the sparse coding approach. We preprocessed the patches following the previous authors’ procedures and also performed additional steps to remove trivial statistical cues. All image processing details are provided in the Methods. Following the previous authors, drawings had been downsampled, so the patches used for stimuli were 16×16 pixels in size. Each texture contained 100 patches and subtended 6 degrees of visual angle on a side. Each observer completed 440 test trials covering all possible combinations of drawings allowed under the three texture task restrictions. All observers discriminated the two categories well above chance levels as shown in Fig. 3, with a range of performance from 67%-83% correct, mean = 73% correct (chance = 50% correct).

----- Fig. 3 about here -----
2.2. Feature Identification Experiments

In Experiments 2 and 3, three of the observers from Experiment 1 returned for two additional sessions, in which we statistically controlled the visual information available in the stimuli. By essentially filtering out some information, we could test whether performance was affected and explore the critical features distinguishing the two categories of drawings.

Experiment 2 was a phase scrambled version of Experiment 1, in which all details of the experiment were identical except that each image patch making up a texture was subjected to Fourier phase randomization. This procedure preserves the Fourier amplitude in each individual patch, but it destroys sharp edges and Gaussianizes the pixel distribution (i.e. a black and white image would have many more tones of gray after phase scrambling). To illustrate these effects, Figure 4A shows the patch-wise phase scrambled version of the stimulus in Fig. 2. Experiment 3 was a Portilla-Simoncelli texture model (2000) matched version of Experiment 1 and was again identical in procedure except that stimuli were not patches from actual drawings, but Portilla-Simoncelli texture model samples with identical texture model features. With this texture model one can create a visual texture by specifying particular values for a set of physiologically inspired features. The particular values of the features we set were those of individual image patches in the drawings. Whereas phase scrambling preserves only the amplitude spectra of each patch, using the Portilla-Simoncelli model to synthesize new patches matches many more textural features of the drawings. An example stimulus is shown in Fig. 4B.

----- Fig. 4 about here -----

Subjects were again significantly above chance in both experiments as shown in Fig. 5, indicating that neither manipulation completely removed the perceptual differences between the two drawing categories that observers used to discriminate them.

Performance dropped significantly for all three subjects from Experiment 1 (mean = 76% correct) to near chance levels in Experiment 2 (mean = 63% correct), indicating that phase scrambling removed many but not all of the identifiable image differences. This means that variations in the dominant orientations of the image structure (e.g. horizontal vs. vertical) and the relative strengths of fine versus coarse scale content differed discriminably between the two categories.

In Experiment 3, only the most sensitive observer performed worse when given only the Portilla-Simoncelli features, but still performed well above chance at 76% correct, indicating that the texture model captured much useful information for discriminating the
authentic from inauthentic drawings. For the other two observers, however, performance was no different with the Portilla-Simoncelli features than with the phase-scrambled images, meaning that these observers could not make use of the additional visual information captured by the texture model beyond fluctuations in local Fourier amplitude spectra.

----- Fig. 5 about here -----

2.3. Performance Analyzed by Drawing

In the previous sections, we reported percent corrects as average performance over all drawings as an estimate of the two categories’ overall discriminability. We also evaluated performance on a drawing-by-drawing basis since the drawings have different characteristics and histories, and it allows clearer comparison with the sparse coding discriminator.

In Fig. 6, we plot the discriminability of each Bruegel drawing from the class of inauthentic drawings after taking both hits and false alarms into account (light gray bars). Hits are trials on which the Bruegel was selected when another Bruegel was the reference. False alarms are trials on which the Bruegel was selected when an inauthentic drawing was the reference. We also plot the percentage of false alarms made for each drawing (white bars). For inauthentic drawings this is the percentage of trials on which the inauthentic drawing was selected although an authentic drawing was the reference. For comparison to the sparse coding discriminator, we plot its performance (from Hughes et al., 2010) with the same size patches (dark gray bars).

In Experiment 1, our observers classified each drawing correctly. In Experiment 2, observers remained significantly above chance for all drawings except no. 9. In Experiment 3, observers again performed significantly above chance with all drawings. Stars indicate where the three observers in Experiments 2 and 3 performed significantly worse than they had in Experiment 1—as assessed by checking whether performance fell outside the corresponding 95% confidence interval in Experiment 1.

In Experiment 2, seven of the eight Bruegel drawings became significantly more difficult to discriminate from the inauthentics when phase scrambling was applied. Only drawing no. 11 was not significantly affected by the manipulation, indicating that the patch-wise Fourier amplitude spectra were sufficient to account for the drawing’s discriminability level in Experiment 1. For all other drawings except no. 9, phase scrambling destroyed most but not all of the visual cues observers relied upon to discriminate the two categories.
In Experiment 3, five of the eight Bruegel drawings became significantly more difficult for the three observers to discriminate from the inauthentics when they were represented as Portilla-Simoncelli textures. The result indicates that for drawings no. 9, 11, and 13, the Portilla-Simoncelli features sufficiently captured the visual information observers used to discriminate the drawings in Experiment 1, but for the remaining drawings, observers relied on additional information as well.

The increased difficulty of Experiments 2 and 3 is reflected by increased alarm rates relative to Experiment 1 (mean false alarm rate = 28% averaged over all drawings and observers) both for Experiment 2 (mean = 41%), \( t(12) = 4.83, p < 0.001 \), and Experiment 3 (mean = 37%), \( t(12) = 3.48, p < 0.005 \).

The dataset contains only one strong authenticity test: the comparison of drawing no. 6 versus no. 7, a direct copy. Of the inauthentic drawings, no. 7 had the highest false alarm rate in all experiments. Whenever no. 6 and 7 were pitted against each other as comparisons (11 trials per observer per experiment), observers were at chance in all three experiments.

3. DISCUSSION

3.1. A New Paradigm to Study Art Perception

We presented a new paradigm for studying how people perceive artistic style. In contrast to previous approaches in which artworks are viewed in their entirety, we propose that viewing a statistical sample of local image patches (e.g. Fig. 1C) can also lead to valuable advances in understanding art perception. Although subject matter and global composition are certainly strong components to an artist’s style, this paradigm focuses viewers on purely perceptual stylistic choices and avoids semantic categorizations—as long as the patches are small enough—which can potentially confound judgments of purely visual qualities. We note that the paradigm is also sensitive to some compositional aspects of style. For example, the overall degree of compositional density will be apparent in a random set of small image patches extracted from the piece. Below we describe the advantages of this new approach.

One important advantage of showing only a few local image patches on each trial is that it allows the experimenter to gather several judgments of the same artwork without ever repeating a stimulus. In our experiments we showed only one hundred 16×16 pixel patches of
an artwork at a time, which corresponded to less than 2% of the whole image. When viewing artworks in their entirety, on the other hand, observers would recognize repeated presentations and could potentially use memory of previous responses to strategize instead of simply judging the visual information present in the stimuli. Running multiple trials with different samples of local image content allows for a more precise estimate of observers’ sensitivity to stylistic features.

The paradigm’s flexibility allows experimenters to measure and manipulate the statistical image features available to observers and therefore to evaluate how people rely on different kinds of visual information to form their judgments. Because a small collection of several local image samples is shown on any trial, the distribution of values taken on by a particular local image feature of interest can be measured or manipulated on a trial-by-trial basis. In our experiments, we were interested in how observers discriminated drawings by different artists, so we used several specialized preprocessing steps (section 4.1.1.) to attenuate irrelevant aspects of the images. We also conducted two further experiments where we filtered out specific image features to measure how they contributed to observers’ judgments, but the paradigm is certainly not restricted to the particular image processing operations we performed. Depending on a particular experiment’s image processing manipulations, we should note that the experimenter can space out the individual image patches slightly (as in Fig. 1C) or tightly tile them together (Fig. 2), which works well when all image patches have the same mean luminance or color value.

In our experiments, the participants were unaware that they were judging artistic style; they simply performed a texture discrimination task (Fig. 2) where stimuli were constructed from two categories of drawings (Table 1). We carefully preprocessed the images, so that judgments could be formed only on the basis of local stroke variations. However, we deliberately described the task to the participants as a texture similarity judgment and not as an artistic style judgment because we were interested only in their perceptual sensitivity to stroke style, not in their preconceived expectations about how artists’ strokes might vary. Our results indicated that even non-experts could reliably discriminate the strokes drawn by Pieter Bruegel the Elder from those of his imitators. In light of previous studies of artistic style where non-experts failed to judge artworks on the basis of stroke style (Augustin & Leder, 2006), our results indicate that the local texture discrimination task is one way to overcome the difficulty of measuring untrained observers’ sensitivity to style. It is a separate question how the textural style qualities captured by our local image patch stimuli contribute to subjective experience, and one which would require future studies with different
instructions. Quantifying observers’ pure perception of textural style qualities—in the absence of semantic expectations evoked by recognizable image content or by knowledge that the task involves art perception—on the other hand, provides an objective measure of visual aspects of style that can be related to mathematical image descriptions.

The paradigm can be adapted to address a broader variety of questions about artistic style perception as well. For example, one could measure aesthetic or other subjective judgments and still maintain use of a forced choice task, which is advantageous since forced choice procedures avoid individual differences due to decision making factors and provide bias-free estimates of discriminability (Green & Swets, 1966). To adapt the task to measure aesthetic judgments, one could present two textures from different artworks on each trial and ask participants to choose which one is more aesthetically pleasing. The results would reveal any potential ranking among the artworks in terms of each observer’s aesthetic preferences for the local, textural image content.

In summary, the main advantages of the new local paradigm are several: 1) semantic categorization of artworks is avoided, 2) a forced choice discrimination task can be used, 3) several trials can be run without repeating a stimulus, 4) identification of critical local image features underlying observers’ judgments is well-defined, and 5) it can be used without observers realizing that they are judging art, in which case individual variation due to idiosyncratic expectations about the definition of artistic style is not an issue.

Artists have no doubt gained a richness of expression by experimenting with contrasts between local and global scales (e.g., pointillism or more recently, Chuck Close’s later portrait style). We suggest that the scientific study of art perception could also benefit from more detailed attention to perception of local scales. It may well reveal new insights into the perception process or even interesting new comparisons of artists, periods, or artworks.

3.2. The Bruegel Drawing Dataset

3.2.1. Sensitivity to Stylistic Differences Detected by Non-Experts

The dataset of drawings by Pieter Bruegel the Elder and his imitators (Table 1) that we used to study artistic style has also been used in several stylometric studies (e.g., Lyu et al., 2004; Hughes et al., 2010 & 2012), in which purely mathematical descriptions of the images were used to classify their authenticity. One particular stylometric study suggested that a sparse coding discriminator can represent meaningful aspects of Bruegel’s stroke style.
We ran a similar classification experiment with non-expert human observers. The results indicated that even non-experts are sensitive to the stylistic differences between the authentic and inauthentic drawings of the dataset, and our most sensitive observer performed at comparable levels to the sparse coding discriminator (Fig. 3).

In two further experiments, we manipulated the local image properties of the drawings in order to examine which mathematically defined image features were used to visually discriminate the two categories of drawings. The results indicated that edge sharpness was an important factor and that image patch Fourier amplitude spectra played a role as well, meaning that local differences in the amount of coarse versus fine spatial structure and/or in the dominant stroke orientations (e.g. horizontal, vertical, oblique) could be used by observers to discriminate the two categories of drawings.

Our non-expert observers were able to discriminate more complex differences in stylistic variations across the drawings as well and therefore did not merely rely on such simple cues as edge sharpness. The group of inauthentic drawings actually comprises two separate stylistic categories: no. 7, a direct copy of Bruegel’s no. 6, versus nos. 120-127 (Metropolitan Museum, 2001). Our results indicate that non-experts are sensitive to this difference. Whenever participants were faced with a trial on which no. 7 and any one of nos. 120-127 were pitted against an authentic drawing, they were at chance (mean = 50% correct). On the other hand, when two drawings from nos. 120-127 were pitted against an authentic drawing, participants were near their top performance levels (mean = 74% correct). This performance difference between no. 7 and the group 120-127 indicates that observers could clearly distinguish the two classes of imitation drawings and were sensitive to the same intricate stylistic distinctions made by experts in art criticism.

3.2.2. Reasons for Caution in Using the Dataset and Interpreting Previous Results with It

The fact that stroke sharpness was an important perceptual cue raises a red flag about the dataset. Certainly, other artists could also pen sharp strokes while having little other stylistic similarity to Pieter Bruegel the Elder, and artistic style discrimination of this sort, i.e. discriminating strokes drawn by one artist versus another, should depend on shape-based stroke features. It is unclear whether stroke sharpness could account for the previously reported stylometric success in classifying the images of this dataset (e.g., Lyu et al., 2004; Hughes et al., 2010 & 2012). One drawing in the dataset suggests this may be the case. Drawing no. 5, one of the authentic drawings, has been extensively reworked by another
artist. Crisp parallel hatching was drawn over Bruegel’s empty sky and foreground, covering about 50% of the drawing in strokes inconsistent with Bruegel’s style (Metropolitan Museum, 2001). None of these regions affected the sparse coding algorithm’s performance, which classified no. 5 as a Bruegel with 100% certainty.

Besides the issue of stroke sharpness, which we note is apparent only in the raw scans of the drawings, not in the catalogue reproductions of the Metropolitan Museum (2001), there are further reasons for caution. First, the imitations have lower contrast than the authentic drawings, even after the extensive preprocessing operations used in the sparse coding study. (In our experiments we removed the remaining differences by globally normalizing each image, and we additionally removed local contrast fluctuation differences as shown in Fig. 8, so neither global nor local contrast cues can explain our participants’ performance.) It is therefore important that potential future studies using this dataset take care to account for contrast cues in reporting their results. Second, although all drawings are landscapes with roughly similar compositional content, the sizes of the drawings vary (Table 1) which affects the scale of image structure present in extracted image patches of a fixed physical size. Such scale differences are potentially perceptible: the surface areas of the authentic drawings (normalized to 1) were marginally correlated with the participants’ performance, $\rho = -0.62, p = 0.10$, such that larger drawings, being the outliers in terms of size (Table 1), were linked with worse performance.

Finally, it is worth noting that the dataset does not contain a set of authentic pieces and direct copies. Instead there is only one such pair in this dataset, and both categories of drawings contain stylistic variations (see Table 1 and the Metropolitan Museum catalogue, 2001). The dataset therefore does not provide the strongest possible test of authenticity judgment, which would be to discriminate compositionally identical drawings penned by different artists.

4. MATERIALS AND METHODS

4.1. General Method

Stimuli were textures made of one hundred square image patches tightly tiled together. The image patches were sampled at a very local scale from drawings in the Bruegel database used in previous statistical analyses of artistic style (Lyu et al., 2004; Rockmore et al., 2006; Hughes et al., 2012). In the digital scans of the drawings, one square centimeter is
represented by approximately $104 \times 104$ pixels. The image patches we used for stimuli were $16 \times 16$ pixels in size and were extracted from grayscale downsampled versions of the digital scans, such that our patches corresponded to $32 \times 32$ pixel regions in the original raw scans or approximately $31 \times 31$ mm regions in the drawings. The database contains two categories of drawings: 8 authentic and 5 inauthentic detailed in Table 1. On each trial, a reference texture and two comparison textures were presented (Fig. 2). Each texture was made of image patches from one drawing only. One of the comparison textures was always from the same category but not the same drawing as the reference texture, while the other comparison was always from the other category. Observers were instructed to select the comparison texture from the same category as the reference. Subjects had unlimited time to respond, feedback was always given, and a short round of one hundred training trials was performed prior to experimental trials. No image patch was repeated at any point during an individual session.

Each subject completed 440 test trials of all possible combinations of drawings within the requirements of the three texture task ($8 \times 5 \times (8-1) + 8 \times 5 \times (5-1) = 440$) and did not necessarily see the same image patches as any other subject. We measured the average percent correct as a measure of the two categories’ discriminability for each subject separately.

4.1.1. Pre-processing of Drawings into Image Patches

The stimuli were prepared after a series of image processing steps had been applied to the raw scanned images of the drawings. This pre-processing comprised two steps: first, replicating the pre-processing used by Hughes et al. (2010), and second, an additional series of pre-processing to remove image features related to mean and contrast fluctuations which human observers find very salient (Gerhard et al., 2013), but which are not related to the style of the strokes drawn. As the patches are sampled from very small regions of the drawings, the paper texture is also apparent. We took further steps to minimize the paper texture’s role in the experiments. We should also note, that we used the entire drawings, not just a single square section of the drawings as Hughes et al. (2010) did. Readers uninterested in the specific details of the pre-processing steps can skip the remainder of this section. We also note that example code is provided on our lab website (http://bethgelab.org/code/).

We first applied the initial pre-processing steps following Hughes et al. (2010) exactly, except where noted. We transformed the color images to grayscale, then resized the images by 50% using cubic interpolation. These images were used as the basis for the Portilla-Simoncelli texture synthesis procedure (2000) in Experiment 3, described in further
The images for Experiments 1 and 2 were subjected to further processing as follows. At this point, Hughes et al. (2010) sampled a random square section from the drawing and discarded the rest from further analyses; however, we wished to retain as much information from the drawings as possible, so we split each drawing into two overlapping square images (square dimensions required for the Fourier filtering step) and later recombined them into a single large image. The square images were whitened according to the previous authors’ filter, which does not completely flatten the amplitude spectrum, but rather enhances higher spatial frequencies relative to lower. It therefore enhanced edges and removed some of the large fluctuations in mean brightness across image regions as shown in Fig. 7. The images were then stitched back together, avoiding edge artifacts. The final step following the previous authors was to rescale the grayscale pixel values of the images linearly to a range from -1 to +1.

At this stage, it is apparent that the inauthentic drawings have lower contrast than the authentic drawings (mean r.m.s. contrast of authentics = 0.14±0.023, of inauthentics = 0.10±0.024), so we z-scored all gray values within each image separately to remove this global image difference between the classes.

Next, to avoid patches that contained only paper texture, we automatically labeled the lowest contrast regions of each drawing separately. To do so, we first sampled 64×64 pixel square image patches at 8 pixel increments horizontally and vertically across the whitened drawing. We measured the standard deviation of the grayscale pixel values within each 64×64 pixel patch and labeled patches in the bottom 20% to be excluded as stimuli. (Although whitened images were not used in Experiment 3, the same labels were used to select patches from the original drawings for processing with the Portilla-Simoncelli algorithm.) Although the proportion of empty space varied visibly across the drawings, we found the 20% threshold to be a reasonable compromise.

Each remaining 64×64 pixel patch was broken into sixteen 16×16 pixel patches to be used in constructing texture stimuli. We removed the mean grayscale value of each 16×16 pixel patch, so that all patches in the experiment would have the same mean gray value, and observers would not be distracted by mean gray level fluctuations across patches and focus instead on the style of the pencil and pen strokes drawn by the artists. Similarly, we removed local contrast fluctuations as a cue in the experiment. This was the only pre-processing step that was applied on a trial-by-trial basis. In our previous work (Gerhard et al., 2013), we found that human observers are highly sensitive to the contrast fluctuations across image
patches in tightly-tiled texture stimuli, so to avoid observers relying on this kind of information, which likely has little to do with the style of the strokes made by the artist, we equalized the histogram over the standard deviations of the 100 patches shown on each trial. The effect of this operation is to have equal numbers of image patches at each contrast level from 0 to 1. Patches which initially had low contrast (likely to be paper texture regions that had survived the earlier threshold) were tuned down to zero contrast, whereas higher contrast patches were tuned upward. An example is shown in Fig. 8.

In summary, the pre-processing steps we applied removed several cues: global image means and contrasts and local fluctuations in mean and contrast across images.

The image patches used as stimuli in Experiment 2 were patch-wise Fourier phase scrambled as a final step before being constructed into textures and were otherwise identical to those of Experiment 1. An example was shown in Fig. 4A.

The image patches in Experiment 3 were synthesized from the grayscale 64×64 pixel patches sampled from the non-whitened drawings. The reason we did not whiten the images was because the Portilla-Simoncelli texture synthesis algorithm is designed to synthesize textures with non-white amplitude spectra. However, we only used the same upper 80% of patches in terms of contrast as in the other experiments, so patch identity in the original drawings was matched across all three experiments. To completely remove contrast fluctuations across patches as a cue, we removed the mean of each 64×64 pixel patch and adjusted the pixel values so the vector norm of the vectorized version of the image patch would equal one. Hence, all patches had the same mean and r.m.s. contrast. The Portilla-Simoncelli texture features of these adjusted patches were analysed and submitted to the iterative texture synthesis procedure of the algorithm using 4 image scales, 4 orientations, a 7 pixel neighborhood, and 50 iterations. The resulting 64×64 pixel image patches retain the same Portilla-Simoncelli features as the input patches, but are not pixel-by-pixel matches. Synthesized patches and originals are shown in Fig. 9. Each synthesized 64×64 pixel patch was broken into sixteen 16×16 pixel patches. Although the original 64×64 pixel patches had the same norm and mean, breaking them into smaller patches introduced mean and contrast fluctuations into the ensemble of 16×16 pixel patches. As in Experiments 1 and 2, we removed all individual patch means and equalized the contrast fluctuations across the 100 patches of each trial individually using histogram equalization.
All of the 16×16 pixel patches in each experiment were magnified by a factor of 2 (returning to the scanned drawings’ original size) before being presented to observers. We magnified the 16×16 pixel image patches to 32×32 pixel patches by simply replacing the pixel value at each location with a 2×2 pixel block of the same value.

4.1.2. Texture Stimuli

Stimuli were constructed by creating tightly-tiled textures for each of the three drawings selected for that particular trial. We first randomly selected 100 random image patches without replacement from the set of image patches (described in section 4.1.1.) for the particular drawing in question. The 100 patches were then tightly tiled into a square texture of image patches. After assembling all three textures, we normalized the range of gray values in the three textures to be between 0 and 1. When presented at the observer’s viewing distance, each texture subtended 6×6 degrees of visual angle. The three textures were evenly spaced horizontally against a black background as shown in Fig. 2.

4.1.3. Apparatus and Software

Subjects performed the experiment in a dark room. Stimuli were displayed on a linearized EIZO 21-inch CRT monochrome display, with maximum luminance set to 563 cd/m². A forehead bar and chinrest fixed viewing distance at 90 cm. We used a custom DATAPixx video controller with 16-bit grayscale resolution and the 5 button RESPONSEPixx response box. The experiment was programmed in MATLAB using the Psychophysics toolbox (Brainard, 1997; Pelli, 1997; Kleiner et al., 2007).

4.1.4. Procedure

Subjects read an instruction sheet prior to entering the experimental apparatus. It showed an example trial and informed them that they would be viewing textures made of small image patches that had been tightly tiled together. They were told that the image patches were random samples cut out of larger images which came from two distinct classes and that their ability to discriminate the classes would be measured in the experiment. They were not told that the images were drawings or even that the study related to art perception. Instead the instructions were presented as a texture discrimination experiment.

Each subject first completed 100 training trials randomly selected from the 440 possible combinations of drawings. Subjects were informed that the first 100 trials would be
practice and that it was their chance to start identifying aspects of the textures that defined the two different classes. There was a clear break from the training to the test trials.

Subjects were allowed to view each trial as long as necessary to make their judgments. Feedback was given after incorrect selections: the textures were shown again with a white frame encircling the correct answer for 3500 milliseconds. During the inter-trial interval of 250 milliseconds, the screen was black.

After every 10 training trials and every 20 test trials, subjects were informed about their performance. Both the percent correct achieved in the previous block and the overall percent current were given. After viewing their results and taking a break if necessary, subjects advanced to the next block of trials by pushing a button.

The order of the 440 combinations of the drawings was randomized for each subject, as were the sides on which the same- and different-class comparison textures appeared. The side corresponding to the correct answer was equally likely to be left or right.

4.1.5. Observers

Seven observers naïve to the purpose of the experiments participated. None of the subjects reported artistic training. Three of the subjects in Experiment 1 returned to complete Experiments 2 and 3 (in separate sessions). All observers had normal or corrected-to-normal vision. Observers gave informed consent prior to participation, were treated in accordance with the Helsinki Declaration, and were compensated 8 Euro/hour for their participation.

5. ACKNOWLEDGEMENTS

We thank Alexander Ecker, David Acquistapace, and the Laboratory of Experimental Psychology at the Katholieke Universiteit Leuven for helpful discussions and Niklas Ludtke for sharing his implementation of the Portilla-Simoncelli texture synthesis procedure. We also thank James Hughes for sharing the dataset of scanned drawings used in previous studies.
6. REFERENCES


<table>
<thead>
<tr>
<th>Catalogue Number</th>
<th>Title</th>
<th>Date</th>
<th>Size (cm)</th>
<th>Artist</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Pastoral Landscape</td>
<td>1552</td>
<td>21.5×31</td>
<td>Pieter Bruegel the Elder</td>
</tr>
<tr>
<td>4</td>
<td>Mountain Landscape with Ridge and Valley</td>
<td>ca. 1552</td>
<td>20.4×29.5</td>
<td>Pieter Bruegel the Elder</td>
</tr>
<tr>
<td>5</td>
<td>Path through a Village</td>
<td>ca. 1552</td>
<td>20.7×33.1</td>
<td>Pieter Bruegel the Elder</td>
</tr>
<tr>
<td>6</td>
<td>Mule Caravan on Hillside</td>
<td>ca. 1552</td>
<td>21.8×30.1</td>
<td>Pieter Bruegel the Elder</td>
</tr>
<tr>
<td>7</td>
<td>Mule Caravan on Hillside</td>
<td>1603</td>
<td>21.7×30.2</td>
<td>unknown</td>
</tr>
<tr>
<td>9</td>
<td>Mountain Landscape with River and Travelers</td>
<td>1553</td>
<td>22.8×33.8</td>
<td>Pieter Bruegel the Elder</td>
</tr>
<tr>
<td>11</td>
<td>Landscape with Saint Jerome</td>
<td>1553</td>
<td>23.2×33.6</td>
<td>Pieter Bruegel the Elder</td>
</tr>
<tr>
<td>13</td>
<td>Italian Landscape (after Domenico Campagnola)</td>
<td>ca. 1554</td>
<td>33.3×46.6</td>
<td>Pieter Bruegel the Elder</td>
</tr>
<tr>
<td>20</td>
<td>Rest on the Flight into Egypt</td>
<td>ca. 1555-1556</td>
<td>20.3×28.2</td>
<td>Pieter Bruegel the Elder</td>
</tr>
<tr>
<td>120</td>
<td>Mountain Landscape with a River, Village, and Castle</td>
<td>late 16th century or early 17th century</td>
<td>35.7×44.4</td>
<td>Master of the Mountain Landscapes</td>
</tr>
<tr>
<td>121</td>
<td>Alpine Landscape (View of Waltensburg)</td>
<td>late 16th century or early 17th century</td>
<td>32×27.4</td>
<td>Master of the Mountain Landscapes</td>
</tr>
<tr>
<td>125</td>
<td>Solicitude Rustica (Rustic Cares)</td>
<td>late 16th century or early 17th century</td>
<td>24.4×35.2</td>
<td>Master of the Mountain Landscapes</td>
</tr>
<tr>
<td>127</td>
<td>Rocky Landscape with Castle and a River</td>
<td>ca. 1590</td>
<td>19.2×31</td>
<td>Jacob Savery</td>
</tr>
</tbody>
</table>

**Table 1. Drawings used for stimuli.** The dataset of landscape drawings studied here and in several stylometric studies contains thirteen drawings: five inauthentic and eight authentic drawings by Pieter Bruegel the Elder (ca. 1525-1569). We use the Metropolitan Museum of Art’s exhibition catalogue number (2001) to refer to each drawing. In the digital scans, one centimeter spanned approximately 104 pixels. Six of the eight authentic Bruegels (no. 3, 4, 5, 9, 11, and 13) are early drawings with repeated trademarks of an Italian influence. One (no. 6) is from the same period and has a similar composition but lacks the characteristic foreground style of tree present in the other six. The eighth (no. 20) demonstrates influences of Hieronymus Bosch although it also retains the earlier Italian style tree. The five inauthentic drawings in the dataset are mixed. One is a copy (no. 7) of a Bruegel drawing (no. 6). Three (no. 120, 121, and 125) sharing stylistic similarities were removed from Bruegel’s oeuvre due to a watermark dating no. 120 at least fifteen years after Bruegel’s death. The three imitate features of a later group of Bruegel drawings than the ones included in the dataset (Metropolitan Musuem, 2001). The fifth inauthentic piece (no. 127) imitates Bruegel’s general style but not his detailed trademark strokes (Metropolitan Musuem, 2001).
Figure 1. Contrasting global and local views. A) The traditional global view. A large portion of an early Pieter Bruegel the Elder drawing (Table 1 no. 11 courtesy National Gallery of Art, Washington). B) A psychophysical stimulus of local samples. Focusing on local regions of an image highlights the textural details. We propose a new paradigm where a random selection of image patches is presented to observers as a statistical sampling of the piece’s fine scale character. Patches can be made small enough to preclude recognition of subject matter. The 64 image patches shown here represent less than 4% of A.
Figure 2. The three texture discrimination task. The task is to select which of the two side textures belongs to the same category as the reference in the middle. Observers were told that two kinds of images would be used in the experiment and that each texture was made of small image patches sampled from a single larger image. Unbeknownst to the observers, the two categories were authentic and inauthentic Bruegel drawings (Table 1). The patches corresponded to regions approximately $31 \times 31$ mm in the drawings. Here, the left texture is made of patches from drawing no. 5 (courtesy Leiden University Libraries, Print Room), the right from no. 7 (courtesy Staatliche Graphische Sammlung München), and the reference is no. 11 (courtesy National Gallery of Art, Washington). The correct answer is therefore “left.” In our experiments, the drawings were preprocessed to replicate the conditions of previous statistical analyses of style (e.g. via whitening) as well as to rule out trivial image cues (e.g. via normalizing and contrast equalization) and focus viewers on the strokes instead. All image processing operations are explained in Section 4.1.1.
**Figure 3. Main results.** In Experiment 1, seven naïve, artistically untrained observers discriminated authentic and inauthentic Pieter Breugel the Elder drawings significantly better than chance (50% correct). Each observer’s percent correct is plotted with the binomial 95% confidence interval, where observers are sorted by overall performance. On average, observers achieved 73% correct (range: 67%-83% correct). The most sensitive observer matched the mean level of performance achieved by a sparse coding discriminator applied to the same dataset (Hughes et al., 2010).
Figure 4. Feature identification experiment stimuli. Left textures are again based on patches from drawing no. 5, right textures from no. 7, and the reference is no. 11. The correct answer is “left.” A) Phase-scrambled image patches. Stimuli were prepared as in Experiment 1 except that each image patch in a texture was also Fourier phase scrambled. These textures can be directly compared to Fig. 2, which shows the unscrambled version. Phase scrambling preserves some image features but destroys edge coherence and changes the distribution of gray values. B) Texture model stimuli. Image patches from the drawings were analyzed using the physiologically-inspired Portilla-Simoncelli texture model (2000), and new images with the same model features were synthesized. Figure 9 also shows a comparison of original versus synthesized image patches. All of the image processing operations are explained in detail in Section 4.1.1.
Figure 5. Feature identification experiment results. Three observers returned for Experiment 2 (phase scrambling) and 3 (Portilla-Simoncelli textures). Their performance in all three experiments is plotted with 95% confidence intervals. The observers always performed significantly better than chance (50%). All observers performed significantly worse in Experiment 2 (mean = 63% correct) compared to Experiment 1 (mean = 76% correct). Only the most sensitive observer, O6, performed worse in Experiment 3 than in Experiment 1.
Figure 6. Discrimination performance by drawing. We plot the discriminability of each authentic drawing from the class of inauthentic drawings (light gray bars) with 95% confidence intervals, taking into account both hits and false alarms. White bars show false alarm rates for each drawing. Stars indicate where the three observers in Experiments 2 and 3 performed significantly worse than in Experiment 1. **A) Experiment 1.** Observers performed significantly above chance with each Bruegel drawing whereas the sparse coding discriminator (dark gray bars) misclassified no. 11 (results from Hughes et al., 2010).  **B) Experiment 2.** Image patches were phase-scrambled. Observers performed significantly better than chance in all cases except for drawing no. 9. Only drawing no. 11 was unaffected by the manipulation.  **C) Experiment 3.** Image patches were Portilla-Simoncelli texture model samples. Observers discriminated each authentic drawing from the set of inauthenticities significantly better than chance. Drawings no. 9, 11, and 13 were unaffected by the manipulation.
Figure 7. **Fourier power spectrum whitening.** As part of the preprocessing steps used for the sparse coding discriminator, each drawing’s power spectrum was manipulated so that higher spatial frequencies (fine details) were amplified relative to lower spatial frequencies (coarse scale variations). This example shows the effects of whitening on a square section of drawing no. 11 (courtesy National Gallery of Art, Washington). For the purposes of illustration, these images’ contrasts have been amplified slightly. Insets show the Fourier power spectra (power versus spatial frequency) plotted on identical log-log axes. Because the drawings were whitened *globally*, local image patch Fourier amplitudes varied and were not necessarily white. **A) Before. B) After.**
Figure 8. Histogram equalization of contrast across image patches. For each texture of 100 image patches, we equalized the distribution over the 100 patch r.m.s. contrast values. The two primary benefits of this procedure are 1) it attenuates the appearance of paper texture, and 2) it removes local contrast fluctuations as a potential cue to category identity. Insets show the histograms over patch contrast on identical axes. (To ease comparison, the contrast values in A were normalized to a maximum of 1 before plotting. In B the maximum contrast equals 1 by definition.) This example shows patches from drawing no. 11 (courtesy National Gallery of Art, Washington). A) Before. B) After.
Figure 9. Original and synthesized image patches from drawings in the dataset. A) Patches from a Bruegel drawing (no. 11, courtesy National Gallery of Art, Washington). B) Patches from an inauthentic drawing (no. 7, courtesy Staatliche Graphische Sammlung München). The upper row shows the original 64×64 pixel patches. Directly below we show the synthesized patches with identical Portilla-Simoncelli features. The model represents information in a much lower dimensional space than the raw pixels and captures several textural qualities of the drawings as shown. Each 64×64 pixel patch was broken into sixteen 16×16 pixel patches for use as stimuli. All details of the image processing operations that were applied are explained in Section 4.1.1.