Structure and Aesthetics in Non-Photorealistic Images

Hua Li* David Mould† Jim Davies‡

Carleton University

ABSTRACT

Non-photorealistic rendering (NPR) has been used to produce stylized images, e.g., in a stippled or painted style. To evaluate NPR algorithms, similarity measurements used in image processing have been employed to assess the quality of rendered images. However, there is no standard objective measurement of stylization quality. In many cases, raw side-by-side comparisons are used to demonstrate improvements in aesthetic quality. This means of comparison often fails to be persuasive due to the small size of demonstrations and the subjective choice of images. We conducted a user study and examined responses of 30 subjects in order to determine two things: whether there exists a relationship between the structural quality and aesthetic quality of non-colored non-photorealistic images; and whether the choice of images matters for side-by-side comparisons.

Our study revealed a statistically significant correlation between the aesthetic and structure ratings given by participants: increases in structural rating coincided with increases in aesthetic rating. Second, participants' ratings of structure and aesthetic were influenced by image content: that is, choice of input images influenced the results of side-by-side comparisons.

Index Terms: I.3.3 [COMPUTER GRAPHICS]: Picture/Image Generation—Display algorithms; H.5.1 [INFORMATION INTERFACES AND PRESENTATION]: Multimedia Information Systems—Evaluation/methodology

1 INTRODUCTION

Non-photorealistic rendering (NPR) [2, 24] has been used to produce stylized images. Existing NPR algorithms provide a wide range of styles, including painting, drawing, illustration, mosaics, and cartoons. Recently, the stylized images by NPR methods have been increasingly and actively employed in computer games, films, advertisements, and websites. NPR researchers recognize that the structural quality of the rendered images affects the aesthetic quality. As a result, many NPR algorithms, including stippling methods [11, 15, 18, 29], screening methods [14, 23], mosaic generation [4, 13, 16], and abstraction [19], emphasize structural awareness in their efforts to improve the quality of stylization. There is a trend towards improving NPR algorithms by introducing structure into algorithmic design. However, there is little understanding of how the aesthetic quality of various algorithms is affected by structure. For this reason, one aim of our study was to investigate whether there is a relationship between structural and aesthetic ratings.

NPR has developed into a mature field over the last two to three decades. However, it is not until relatively recently that NPR researchers have made systematic efforts to evaluate and validate NPR algorithms. Comparisons of performance based on processing speed are ill-suited to answering questions such as: "Does the system provide critical effects for this style?", "Do different styles have any differences in visual perception?", or "Do the resulting styles look more appealing than the other resulting styles?". Side-by-side comparisons between stylized images from different NPR algorithms have become a common way to evaluate various NPR approaches. However, as NPR algorithms have become more mature, improvements become smaller; subtle differences in quality are difficult to detect by raw comparison. Computer vision contains objective metrics for quality of filtered images in terms of tone similarity [30], structure similarity [32], and visual appearance [31]. These objective measurements have been used directly in the field of NPR to evaluate stylized images, despite the fact that researchers are uncertain whether these metrics are appropriate for NPR evaluation. To the best of our knowledge, there do not exist metrics for quality of stylization. The goal of this paper is to conduct a user experiment to evaluate various stylization effects. We also hope to determine the effectiveness of side-by-side comparisons and objective measurement.

In order to avoid the influence of color in our study, the images were rendered in black and white or grayscale. The main finding of our user study is the positive correlation between structural and aesthetic ratings, which indicates that participants thought that images with better structural scores are also more visually appealing. This correlation suggests that designers of stylization algorithms may need to consider structural aspects in terms of clarity of specific objects. Our second main finding is that participants' ratings of structure and aesthetics are influenced by image content. In our study, participants assigned higher ratings to images from the Bird category than to those from the Person category in terms of structural and aesthetic quality; participants also responded faster to the Bird category than to the Person category. In addition, image category had a significant effect on ratings and the ratings for NPR algorithms changed depending on image category, which means the choice of image content matters in comparisons.

2 PREVIOUS WORK

Early research in halftoning has mainly focused on improving quality in terms of tone similarity. Floyd-Steinberg error diffusion [5] and Ostromoukhov's halftoning method [21] used error diffusion to maintain tone; to generate a stippling style, Secord's stippling method [27] employed tone matching to simulate the tone of a reference image; screening [31] uses a set of dithered patterns based on tone to approximate a continuous tone in black and white. However, researchers were not satisfied with the stylization quality from tone similarity and tried to include structural awareness into halftoning [1, 12, 22], stippling [11, 15, 18, 29], screening [14, 23], mosaics [4, 13, 16], and abstraction [19] so as to maintain the detail and structure of the input image. Interest in structural quality is a trend in current NPR research. NPR researchers have adopted objective measures borrowed from computer vision to evaluate tone similarity (peak signal-to-noise ratio) [30] and structural matching (the mean structural similarity measure) [32] of stylized images compared to the original. The study of NPR, however, is not focused on similarity to the reference image, but is concerned with visual aesthetics. The influence of structure quality on stylization generation requires further research. One of our main motivations in this paper is to investigate the relationship between structure and aes-

*e-mail:hua.li.qc@gmail.com

†e-mail:mould@scs.carleton.ca

‡e-mail:jim@jimdavies.org

thetics of stylized images. This exploration was done by analyzing user responses.

To evaluate NPR algorithms, researchers often use side-by-side raw comparisons. Hertzmann [8] discussed evaluating human aesthetics and how people respond to stylized images, which provided numerous connections between perceptual evaluation and NPR research. Isenberg's chapter on NPR evaluation [24] classified the evaluation of NPR algorithms into quantitative and qualitative methods.

2.1 Quantitative Evaluation

Different studies used different measurements based on the various applications of NPR. Three types of data can be found in perceptual evaluations for NPR research: scores of study content [17], response time based on recognition or memorization [7, 33, 34], and eye-tracking data [3, 20, 25].

Subjective rating scores in terms of characteristics have been widely used in evaluating NPR algorithms. Schumann et al. [26] conducted early work in assessing the effect of non-photorealistically rendered images by scoring participants' responses on a 5-point Likert scale. Schumann et al. found that that the stylized sketches received significantly higher ratings than the original photos. Gooch and Willemsen [6] were the first to introduce a non-realistic virtual environment to study virtual distance perception. The immersive study provided a way to understand the effectiveness of NPR in a VR environment. Recently, Mandryk et al. [17] and Mould et al. [20] studied the effect of NPR imagery on emotions by collecting rating scores along the dimensions valence, arousal, dominance, and aesthetics. They found that stylized algorithms dampened participants' responses in terms of arousal (activation) and valence (pleasure). Related to affective reactions, Seifi et al. [28] designed color palettes to look at emotional responses to non-photorealistic portraits and reported that when the palette matches the facial expression, the perceived emotion is emphasized; non-matching palettes, however, dampen the perceived emotion.

The second quantitative means of evaluating NPR images is through response times. This method assumes that better quality images will shorten the user response time in a recognition task. Gooch et al. showed that illustrations of faces were more effective than photographs of faces for facial recognition [7]. However, participants recognized caricatures slower than photographs. Winnemöller [33] conducted a similar experiment concerning the ability to recognize and memorize objects using abstractions of arbitrary images. In general, his findings were similar to those of Gooch et al., finding that recognition was significantly faster with the abstracted images as opposed to real photographs.

The third means is the use of eye-tracking devices, which presents an objective measure of cognition. Mandryk et al. [17] and Mould et al. [20] employed eye-tracking and summarized the users attention to stylized images with heat maps. They reported that the stylized images sometimes contained confusing elements or lacked important details, making interpretation difficult.

2.2 Qualitative Evaluation

Schumann et al. [26] evaluated sketched styles by a questionnaire-based approach to obtain qualitative feedback. Isenberg et al. [9] employed a qualitative approach to find out how participants think about hand-drawn illustrations vs. computer-generated output. Mandryk et al. [17] and Mould et al. [20] had users rank eight chosen algorithms based on preference in a study of emotional response and visual attention to non-photorealistic images.

In the current study, we use qualitative evaluation and quantitative evaluation to evaluate the rendered images from stylization algorithms. Participants completed a recognition task and the response time is recorded. Participants also rated scores of aesthetic and structural quality, and finally, ranked all styles. Unlike previous studies, the images used in our study were black and white or grayscale. Because the influence of color was removed from our study, participants had to rely on the organization of primitives (pixels, dots, lines, and curves) to determine the quality of structure and aesthetics. We also analyzed the effects of image content category on ratings and hoped to discover the interaction between image category and NPR algorithms, which have been ignored in previous NPR experimental studies.

3 User Experiment

3.1 Generating Stimuli

The study aims to collect participants' ratings on the quality of rendered images. The images used in the study were downloaded from Flickr.com and were creative commons-licensed for adaptation and modification. We tried to choose images in which the subjects were equally clear and pleasing, so that any differences in the user ratings would arise from the differences in stylization algorithms. Note that since the input images do differ, this remains a possible confound. The images were grouped into seven categories: Car, Cat, Bird, Person, Mug, Flower, and Building. Each category represented a type of object that would be familiar to participants. We used high-resolution images that contained a prominent object from one of the seven categories. After the images were selected, they were converted to a resolution of 828 by 621 and stored as PNG files. We rendered the image stimuli in several styles using existing algorithms from the NPR literature. After we configured parameters, each source image was automatically rendered by the algorithms into black and white or grayscale abstracted styles. Each category contained 13 different images. One of the 13 images was unprocessed; the other 12 images were rendered by different stylization algorithms. The following lists the 12 algorithms used for rendering the images:

- Six structure-aware NPR algorithms
    1. structure-preserving stippling (SPS) [15]
    2. content-sensitive screening (CSS) [14]
    3. hatching with exclusion-based masks (SPH) [15]
    4. artistic tessellation (AT) [13]
    5. line drawing from stippling (Drawing)
    6. line drawing from thresholding edge tangent field (ETF) of Kang et al. [10]
- A contrast-aware halftoning algorithm (CAH) [12]
- Two tone-based NPR algorithms
    1. weighted Voronoi stippling (Secord) [27]
    2. mosaics (Mmosaics) using distance in a weighted graph by Mould [18]
- Black and white stylization (BW)
- Two algorithms that strictly reduced the image information
    1. blurring (Blurring)
    2. adding salt-and-pepper noisy (Noisy)

The stylized images cover a range of styles with different primitives: halftoning (pixels), screening (pixels), stippling (dots), line art (lines), mosaics (regions), and black and white (regions). Examples of the experimental images are shown in Figure 1.

For the generation of CAH [12], SPS [15], CSS [14], Secord [27], and ETF [10], we used exactly the methods described

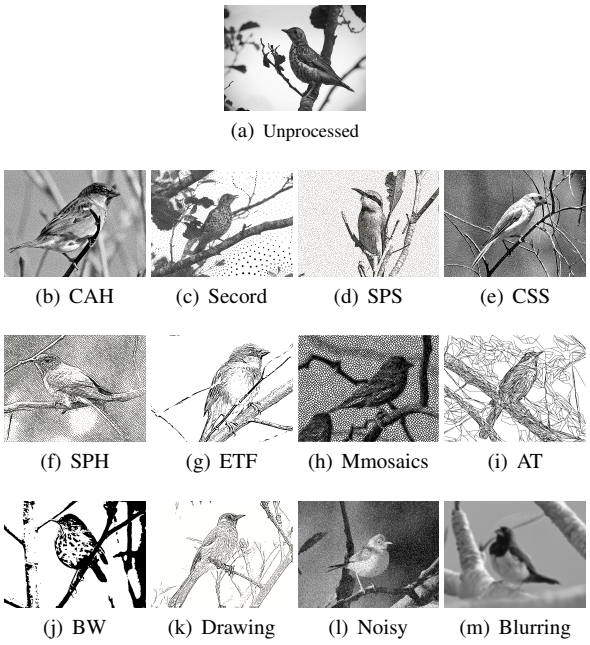


Figure 1: The experimental images from the Bird category. (a) unprocessed; (b) halftoning; (c) stippling; (d) stippling; (e) screening; (f) hatching; (g) ETF; (h) mosaics; (i) mosaics; (j) black and white; (k) drawing; (l) noisy; (m) blurred.

in the original papers. The hatching with exclusion-based masks (SPH) was a variant of sparse SPS stippling using 135° exclusion-based masks to display a hatching style with only diagonal lines. The drawing was converted from a sparse SPS stippling by replacing dots with short strokes. The lines on the edges used the ETF vector field for stroke direction; the directions of lines not lying along the edges were rotated 90 degrees. The AT was tessellation by curves growing from the stippling. The black and white style (BW) was generated by reversing the error sign when doing error diffusion in CAH halftoning [12]. The mosaics (Mmosaics) were generated by first creating stippling using the technique of Mould [18]; each stipple then forms the centre of a region whose size roughly reflects the local image intensity. The noisy images (Noisy) were generated by adding 50% salt and pepper noise to the original photos. The blurred images (Blurring) resulted from processing the originals with a Gaussian filter, using $\sigma = 2.0$.

3.2 Task

The task consisted of participants rating their responses to the images. Participants rated both the aesthetic quality of each image and the perceived clarity of a specified object within each image; the latter rating we refer to as "structural quality". User response times were also recorded. At the end of the study, participants ranked the twelve stylized images. Prior to the experiment, participants completed a training task with a small number of images, using the same training software as the formal study; doing this offered participants a chance to familiarize themselves with the operation of the study. Participants were also encouraged at this time to ask questions related to the study and software.

Following the completion of the training, participants proceeded with the formal study, beginning with the recognition task. Each category starts with a piece of text indicating the search category on the screen and then is followed by a set of rendered images. For example, when participants begin the Car category, the system first shows the text 'Car' on the screen and then the next screen shows a rendered image. This rendered image might or might not include a car object. Participants were asked to press 'A' using the keyboard, indicating 'YES' to accept the category, or to press 'D', indicating 'NO', to reject the category. Participants were asked to respond as quickly as possible. Even if they made a mistake, they could not change their answer. They had to press the 'Enter' key to go to the next image. The interface used for the response time is illustrated in Figure 2.

In the recognition task, the experimental images in each category were shown in a random order. The response time for each rendered image was measured from the moment the image was shown until the participant submitted a response by typing 'A' or 'D' on the keyboard to indicate 'YES' or 'NO'.

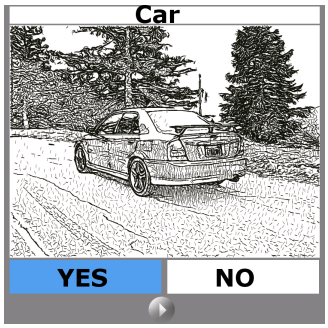
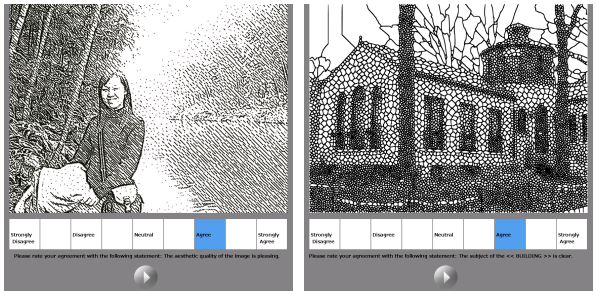


Figure 2: The interface for collecting the response time. The interface contains an experimental image of Drawing for the Car category.

In the second phase of the study, participants rated their responses to the rendered images. The task of rating the aesthetic quality was first and was followed by rating the structural quality. Both aesthetic quality and structural quality were rated using a 5-point Likert scale: Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree. Here, an image was shown on the screen accompanied by a rating scale. After rating the image, participants pressed the submission button when ready to move on to the next trial. Examples of the interface for aesthetic rating and structural rating are illustrated in Figure 3. The aesthetic rating was determined by the participant's response to the image when presented with the prompt: "Please rate your agreement with the following statement: The aesthetic quality of this image is pleasing." In Figure 3, the structural rating was based on the participant's response to the image when given the prompt: "Please rate your agreement with the following statement: The subject of the <<Category Name>> is clear." The "category name" placeholder in the literal text was replaced by the name of the category the image belonged to. Each image includes only one major object of a category surrounded by an arbitrary background. Participants are asked to rate the clarity of the specific major object; we refer to their ratings as "structural quality". Under the assumption that all original images are equally clear and pleasing, differences in subjective ratings are due to differences in the stylization processes.

After participants rated all images, we asked them to fill out a post-experiment questionnaire online. The questionnaire collected the participants' basic information and their preferences of artistic styles. We also asked participants to provide freeform comments regarding possible improvements to the study procedure and enhancements to the quality of stylized images. The final task for participants was to rank the 12 rendered styles (not including the unprocessed photos) based on their preferences. The entire experiment took between 1 and 1.5 hours to complete and participants were given $15 to thank them for their participation. The ethics board at Carleton University approved the experimental protocol.



(a) Aesthetic rating (b) Structure rating

Figure 3: The interfaces for rating the aesthetics and the structure. The left interface contains an experimental image of SPH for the Person category and the right interface contains an experimental image of Mmosaics for the Building category.

3.3 Apparatus

We conducted the experiment with a Windows Vista computer and a 21 TFT display. The resolution of the display was 1440 by 900. The experimental software was coded in Processing. All images were presented at a resolution of 828 by 621 pixels. The system interface displayed images, ratings, and operations on the screen. Participants used the keyboard or the mouse to enter their responses.

3.4 Participants

We recruited 30 participants, aged 18 to 33, of which 15 were female. Participants all had normal or corrected-to-normal vision and did not have any color vision deficiencies. If participants have been practicing artistic work for more than two years, they are referred to as artists. Eleven participants were artists.

3.5 Analysis

The response times of participants varied widely. Participants might have been distracted or might not have responded to the stimuli as quickly as possible, which made the some of the recorded response times unusually long. We suspected that the longest measured times might be problematic, and hence we removed outliers by sorting according to time and then removing the top 10% of the values. The z-value $z = \frac{t-\mu}{\sigma}$ was then used to calculate standardized variables for the remaining response times. The value of z measures the distance between the time t and the mean μ in units of the standard deviation σ. Most participants made 0-2 errors over almost 100 trials, with an overall error rate of about 2%.

We employed Spearman's rank correlation coefficient (*rho*) to assess the correlation between structure and aesthetics. The null hypothesis is: $H0$: There is no correlation between structural and aesthetic ratings.

We used analysis of variance tests (ANOVAs) to determine whether or not there were any statistically significant results in our study. Here, we used an alpha level of 0.05 for all tests. If there was a significant difference, pairwise comparisons of significant results used the Bonferonni method of correcting for multiple tests. We conducted ANOVAs for dot-based methods, region-based methods, the effect of algorithms on ratings and response time, the effect of categories on ratings and response time, and the interaction between image categories and algorithms.

4 Results of Correlation Tests

Are there any correlations between aesthetic ratings and structure ratings? Yes, the correlations are significant ($p \approx 0.000$) after excluding the ratings for unprocessed images. The correlation test between aesthetic and structural ratings showed Spearman rank-order was $rho = 0.279$ (see Figure 4). Excluding the ratings for noisy and blurred images does not change the correlation results much ($p \approx 0.000$, $rho = 0.299$), suggesting that this effect can be attributed to the image filtering methods. This indicated a positive relationship between the ranks obtained in the Structure-Aesthetic rating. Hence we can reject the null hypothesis that there is no correlation between Structure-Aesthetics ratings: in our data the structural ratings were correlated with the aesthetic ratings. Statistically, higher structural ratings accompanied higher aesthetic ratings, which showed that participants associated clearer objects in rendered images with a more appealing visual appearance overall.

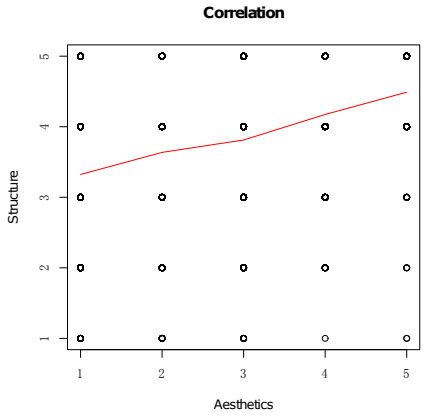


Figure 4: Positive correlation between structure and aesthetic ratings.

There was a significant negative, though weak, relationship between structural rating and response time ($rho = -0.071, p = 0.001$). There was also a significant negative and weak relationship between aesthetic rating and response time ($rho = -0.047, p = 0.036$). Because these relationships are so weak, we cannot conclude anything based on the response times.

There was no significant correlation between response time and MSSIM measurement of structure matching ($rho = 0.019, p = 0.415$). There was no significant correlation between Structure ratings and MSSIM ($rho = 0.015, p = 0.482$). These findings indicated that although the objective MSSIM measure is useful for measuring the overall structural fidelity of an image compared to the original, it does not predict whether images would be easily understood by viewers. There was a significant negative and weak relationship between aesthetic ratings and objective MSSIM measures for rendered images ($rho = -0.114, p \approx 0.000$). MSSIM measures the overall matching between a stylized image and a reference image. However, more structural content might include more irrelevant information. By distracting viewers, irrelevant content might make the overall visual appearance less pleasing.

There was a significant negative and weak relationship between aesthetic ratings and PSNR measurement of tone matching ($rho = -0.118, p \approx 0.000$), which is consistent with the relationship between aesthetic rating and MSSIM. There was a significant positive and weak relationship between structural ratings and PSNR ($rho = 0.059, p = 0.007$). There was no significant correlation between response time and PSNR ($rho = -0.022, p = 0.344$). Usually, the PSNR measurement is not as robust as the MSSIM measurement.

5 Results of Analysis of Variance Tests and Pairwise Comparisons

In the remaining section of this paper, we report on the results for tests with a statistically significant difference. Insignificant results

will not be mentioned.

5.1 Dot-based Methods and Region-Based Methods

Are there overall differences between previous tone-based method (Secord) and the structure-preserving stippling method (SPS)? Dot-based methods consist of Secord's stippling method and the SPS method. The results of ANOVAs indicated that there was a statistically significant difference between Secord's method and the SPS method for both aesthetic rating ($F = 25.507, p \approx 0.000$) and structural ratings ($F = 26.239, p \approx 0.000$). Given that the omnibus test was significant, it was safe to continue with pairwise comparisons. The results of the pairwise comparisons indicated that there was a statistically significant difference between Secord's method and the SPS method ($p \approx 0.000$). The means for the aesthetic and structural ratings are shown in Figure 5. The aesthetic means are 2.7 for Secord and 3.2 for SPS; the structure means are 3.7 for Secord and 4.2 for SPS. Based on these means, we were inclined to conclude that the SPS method is superior to the Secord's method based on aesthetic ratings and structure ratings. Because Secord was only concerned with tone matching in his algorithm, while SPS stippling was focused on structure, the SPS stippling presented clearer object silhouettes and thus the stippled images produced by SPS were preferred. There was no significant difference on the response time based on ANOVAs ($F = 0.002, p = 0.964$).

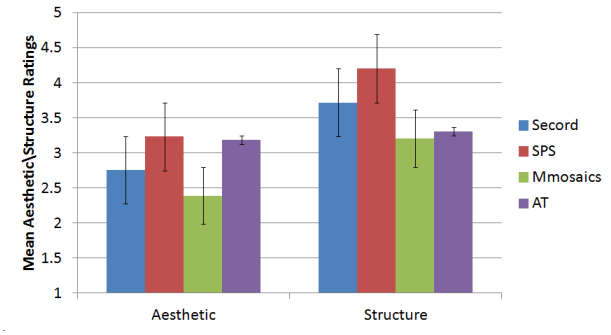


Figure 5: Means +/- SE for aesthetic and structure ratings by the aesthetic or structure of the image stimuli for dot-based methods and region-based methods.

Are there overall differences between tone-based method (Mmosaics) and the structure-aware mosaic method (AT)? As for region-based methods, the one-way ANOVA for the aesthetic ratings ($F = 51.236, p \approx 0.000$) indicated that there was a significant difference between them; however, there was no significant difference in the structural ratings ($F = 0.909, p = 0.341$) or the response time ($F = 0.910, p = 0.341$). Further pairwise comparisons for aesthetic ratings showed there was a statistically significant difference between the Mmosaics method and the AT method ($p \approx 0.000$). Although there was insufficient statistical support to determine that the structural quality of AT was superior to Mmosaics, users thought the AT tessellations (mean=3.1) look more aesthetic than Mmosaics (mean=2.3).

Secord and Mmosaics had lower structural and aesthetic ratings than structure-aware methods such as SPS and AT and we did not list Mmosaics and Secord in the remaining tests. Participants can better understand the specific objects due to clearer expressions of the major structure. This finding of structure-aware methods having higher ratings than tone-based methods also partially demonstrated that current research trend on structure preservation in the NPR field is reasonable. One suggestion from this finding is that containing more edges in stylized images can be a good way for structural improvement.

5.2 Other Results of ANOVAs

5.2.1 Effect of Algorithm on Ratings

Are there overall differences in the user responses to the various NPR algorithms? Using the overall ANOVA, we found a significant effect of algorithm on the aesthetic rating ($F = 55.948, p \approx 0.000$), structure rating ($F = 38.423, p \approx 0.000$) and response time ($F = 2.635, p = 0.003$).

For the aesthetic ratings, after pairwise comparisons, the unprocessed image ($p \approx 0.000$) and the blurred image (Blurring) ($p \approx 0.000$) were significantly different from other rendered images. Figure 6 showed images rendered by CAH, CSS, SPS, SPH, ETF, AT, BW, Drawing, and Unprocessed had higher aesthetic quality than noisy and blurred images. The ETF aesthetic ratings were statistically significantly different from the aesthetic ratings of CAH ($p = 0.046$), and marginally significantly different from CSS ($p = 0.054$) and Drawing ($p = 0.064$); ETF had lower ratings. The CAH and CSS methods produced higher quality aesthetic images, possibly because these methods sought both structure and tone matching; the ETF method is a stylized edge detector without tone matching. Although the ETF style looked more simplified than the Drawing images, the ratings for Drawing were better than the ratings for the ETF in this test.

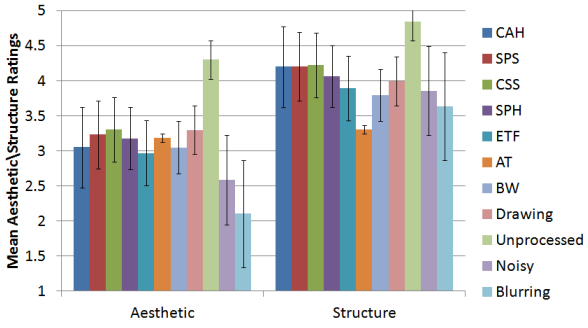


Figure 6: Means +/- SE for aesthetic and structure ratings by the aesthetic or structure of the image stimuli for 10 rendered methods and unprocessed images.

For the structure rating, after pairwise comparisons, the unprocessed image ($p \approx 0.000$) and the AT image ($p < 0.011$) were significantly different in participants' responses from other methods. In this test, although the AT image had lower structural quality than the images rendered by other algorithms, participants thought the AT image was more aesthetically pleasing than the noisy image and the blurred image. The BW structure ratings were statistically significantly different from the structure ratings of CAH ($p \approx 0.000$), CSS ($p \approx 0.000$), and SPS ($p \approx 0.000$) with a lower mean structure ratings than CAH, CSS, and SPS. The CAH structure ratings were statistically significantly different from those of the ETF ($p = 0.0299$) and noisy images ($p = 0.007$). Generally, the CAH image had higher scores in structural quality than the ETF image and the noisy image. It seems that the ETF strategy, always using important edges to represent an image, might be better. However, because the ETF image conveyed no information about tone and CAH is concerned with both structural and tone matching, the CAH image appeared clearer. In addition, the structure ratings of SPS image were statistically significantly different from the structure ratings of CAH, ETF, AT, Unprocessed, Noisy, Blurring (all at $p < 0.025$). Figure 6 showed the SPS had equal to or higher

quality than other algorithms in this structure test, except that the Unprocessed image was rated the highest. Since CAH and SPS are based on the same scheme, we might expect them to receive similar scores. The SPS image had both tone and structure similarity to the original image while the ETF image had no tone matching. The AT image had no clear contrast between edges.

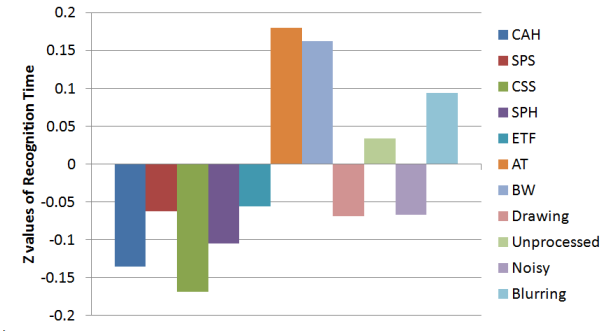
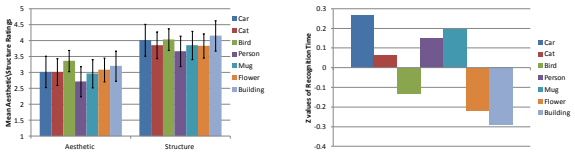


Figure 7: Means for response time.

Figure 7 illustrates the means for recorded response time. The CSS images produced faster response times than AT images at $p = 0.046$, after pairwise comparisons. This finding may indicate that the content of the CSS image was more easily recognized than that of the AT image. This finding is consistent with previous structural rating where the mean CSS structure ratings was higher than the mean AT structure ratings.

5.2.2 Effect of Category on Ratings

Using the overall RM-MANOVA, we found a significant effect on category through the aesthetic rating ($F = 8.509, p \approx 0.000$), structure rating ($F = 7.089, p \approx 0.000$) and response time ($F = 17.024, p \approx 0.000$). Figure 8 illustrates the means for aesthetic and structure ratings by the aesthetic or structure of the image stimuli as well as the means for response time, for the 7 categories.



(a) Aesthetic& Structure Ratings (b) Response Time

Figure 8: (a) Means +/- SE for aesthetic and structure ratings by the aesthetic or structure of the image stimuli on 7 categories; (b) Means for response time on 7 categories.

The aesthetic ratings of the Bird category were statistically significantly different from other categories except the Building category (the Flower category marginally at $p = 0.057$). The Person category was statistically significantly different from other categories except the Mug category (the Car category marginally at $p = 0.064$ and the Cat category marginally at $p = 0.051$). Generally, the Bird category had the highest aesthetic ratings and the Person category had the lowest aesthetic ratings; the Building category also had high ratings, the Mug category had generally low ratings, and the Car, Cat, and Flower categories were intermediate. We suggest that objects with clear silhouettes and substantial interior detail are most likely to be assessed highly; the NPR algorithms will tend to show strong edges such as silhouettes and to place primitives in interior portions of the object, so objects such as birds and buildings with texture (feathers) or interior details (architectural elements such as windows) will be treated well. Objects lacking clear silhouettes (cat, because of fur) or with smooth gradients rather than texture (cars, mugs, many flowers) will tend to look worse. The Person category is especially problematic because human viewers are sensitive to nuances in images of people and will assess each detail carefully, whereas the algorithms lacked semantic knowledge of the image.

The structure ratings of the Person category were statistically significantly different from the Bird category (at $p = 0.003$), the Building category (at $p \approx 0.000$), and the Car category (at $p = 0.006$). The structural quality assessed for the Person category generally was lower than the Bird, Building, and Car categories; in general, people have higher standards for evaluating images of other people than for images of other objects.

For response time, after pairwise comparisons, the response times of the Building category was statistically significantly different from other categories (at $p \approx 0.000$) except the Flower category. It took the shortest time to recognize the Building category. The response time of the Flower category were statistically significantly different from other categories (at $p \approx 0.000$) except the Bird and Building categories. It took the second shortest time to recognize the Flower category. The response times of the Car category were statistically significantly different from other categories (at $p \approx 0.000$) except the the Cat, Person and Mug categories. It took the longest time to recognize the Car category. It seems it was easier for participants to read natural scenes or animals (Building, Flower, Bird, Cat) since the means of response time was as follows: $Building < Flower < Bird < Cat < Person < Mug < Car$. Man-made objects with smooth surfaces such as Car and Mug were difficult to read in an image for participants. The reason might be that people have a clearer outline in mind for smoothed man-made objects than for natural scenes or animals and a little distraction to smoothed man-made objects might affect the visual response. Another reason might be that the algorithms we tested have difficulty portraying smooth surfaces.

5.3 Interaction between Image Category and Algorithm

Did the ratings for the different algorithms change depending on the image category? In addition to showing that the image category was yielding consistent aesthetic ratings, the aesthetic ANOVA also showed that there was a significant interaction between image category and algorithm on the aesthetic ratings ratings ($F = 1.330, p = 0.047$). As Figure 9 shows, for the 7 categories, there was a large difference in the aesthetic ratings. Figure 10 shows the same data as Figure 9, but organized differently. For pixel-based CAH, ETF line art, SPH, and Noisy, participants gave the highest scores for aesthetic quality to the Bird category and the lowest scores to the Person category. For pixel-based CSS, participants rated the highest scores for aesthetic quality to the Bird category and the lowest scores to the Flower category. For dot-based SPS, participants assigned the highest scores for aesthetic quality to the Bird and Flower category and the lowest scores to the Person category. AT and BW got the highest aesthetic ratings in the Building category but got the lowest aesthetic ratings in the Person category. Drawing got the the highest aesthetic ratings in the Car and Bird category but the lowest aesthetic ratings in the Flower category. Blurring got the highest aesthetic ratings in the Flower category but got the lowest aesthetic ratings in the Building category. The structural ANOVA showed that there was no significant interaction between image category and the algorithm in structural ratings (at $p = 0.223$).

The response ANOVA revealed that there is a significant interaction between image category and algorithm on the response time ($F = 2.120, p \approx 0.000$). It is statistically supported that there were four categories: Car ($p = 0.005$), Bird ($p = 0.024$), Mug

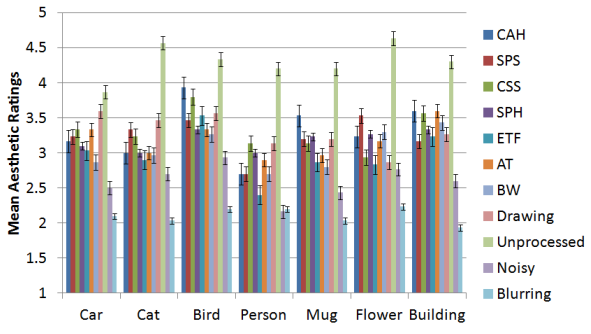


Figure 9: Means +/- SE for aesthetic ratings by the aesthetic of the image stimuli for 7 categories.

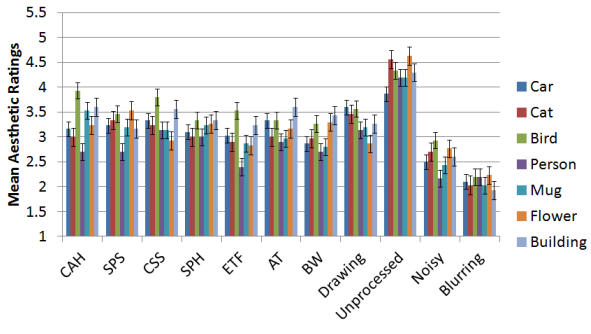


Figure 10: Means +/- SE for aesthetic ratings by the aesthetic of the image stimuli for 7 categories. This is the same as previous Figure only with different visualization.

($p \approx 0.000$), Flower (marginally at $p = 0.061$), where there was a large difference in the response time for the different algorithms. Figure 11 surprisingly indicated that participants recognized the

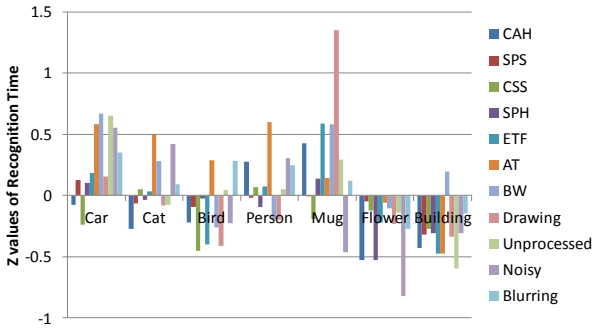


Figure 11: Response time.

images rendered by reduction algorithms including Blurring and Noisy faster than other stylized images by NPR algorithms. Especially for the Mug and Flower categories, the Noisy images can be recognized the most quickly. A possible reason could be that the images for Noisy and Blurring algorithms were grayscale, while the other stylization algorithms used only black and white. The colors provide a lot of information, which ameliorated the information loss caused by blurring or noise. The original image was located in the middle of the order for each category, suggesting that some rendering algorithms improved the response time but some rendering algorithms delayed the recognition time.

6 Minor Findings

Artist and Non-Artist: There was no significant difference between artist and non-artist response times. However, there was a significant effect between artist and non-artist on aesthetic ratings ($p \approx 0.000$), and further pairwise comparisons ($p \approx 0.000$) found that artists (aesMean = 3.42) rated higher than non-artists (aesMean = 3.19) overall. There was a significant difference between artist and non-artist on structure rating ($p \approx 0.000$) and further pairwise comparisons ($p \approx 0.000$). Artists (strMean = 4.47) rated higher than non-artists (strMean = 4.05). One possible explanation is that artists are more familiar with different visual styles and may be more willing to accept varied styles with unusual characteristics.

Female and Male: There was a significant difference between females (aesMean = 3.48) and males (aesMean = 3.07) on aesthetic rating ($p \approx 0.000$), but pairwise comparisons did not provide enough support ($p = 0.17$). There was no significant effect between female and male on structure rating. There is no significant difference between female and male on response time. These findings indicated that, in general, gender did not affect the ratings and the response time.

Overall Ranks after Study: We asked participants in a post-experiment questionnaire to rank the 12 styles rendered by the chosen algorithms used in the study from their most favourite to the least-favourite style. Participants preferred the AT images (7/30 responses) the most, and CAH second (6/30). Participants' least favourite were the blurred images most often (20/30 responses), and with AT second (5/30). Surprisingly, the AT mosaics were very well liked by some participants, but also strongly disliked by others. We suspected that AT might work well for highly-textured regions but might fail for smooth or uniform regions. Thus, images rendered using AT have inconsistent quality in terms of structure and aesthetics. The mixture of quality is favored by some people, but disliked by others.

7 Conclusions and Future Work

This study explored user responses to computer-generated non-photorealistic images. We conducted a 30-subject study measuring aesthetics, structure, and response time over a set of 7 image categories shown in 13 different styles: 12 stylization algorithms and the original image.

The most general finding from the user experiment was that there was a positive correlation between structural and aesthetic ratings. To the extent that there is a causal link from structure to aesthetics (speculated but not known), this provides advice for algorithm designers to consider structure as a possible way to increase aesthetic appeal.

The SPS dot-based method and AT region-based method produced higher quality images than the previous Secord method and Mmosaics region-based method. This is because SPS and AT consider structure preservation more than strictly tone-based methods. This finding further supports our suggestion that structure may be a factor in obtaining aesthetic quality.

Generally speaking, the effect of category on aesthetic ratings and structural ratings showed that Bird images were the easiest images to abstract, while Person images were the most difficult. Our study indicated that NPR researchers, in evaluating the quality in terms of structure and aesthetics, must also consider the choice of the images used. The aesthetic ratings and the response time for the different algorithms change depending on the image category.

The other findings of the study are summarized as follows:

- The rendered images produced significant differences in structural and aesthetic ratings, showing that the experimental stimuli were effective;
- Images produced using Noisy and Blurring are least pleasing and had lowest structural ratings. Images produced using ETF were less aesthetically pleasing than images produced using CAH; images produced using AT were less clear than images produced using CAH, CSS, SPS, ETF, SPH, BW, and Drawing;
- Participants recognized the objects in the images produced by CSS faster than in the images from AT;
- Images in the Bird category had higher aesthetic ratings than images from the Person, Mug, Cat, and Car categories. The Person category had lower aesthetic ratings than did the Flower, Building, and Bird categories. The Person category had lower structural ratings than the Bird, Building, and Car categories;
- Artists generally assigned higher scores than non-artists for both structural and aesthetic ratings;
- Participants preferred the AT and CAH images over all other algorithms and least liked the blurred and AT images. The controversial ranking results for AT are interesting.

In the future, we are planning to include not only more participants in our study, but also more categories of images in order to better evaluate stylization algorithms.

ACKNOWLEDGEMENTS

We would like to thank all participants for spending their time on this study. We also would like to thank Flickr.com for providing images. Many thanks to Sarah Thorne for suggesting revisions on a draft of this paper. This work received financial support from GRAND, NSERC, and Carleton University.

REFERENCES

[1] J. Chang, B. Alain, and V. Ostromoukhov. Structure-aware error diffusion. *ACM Trans. Graph.*, 28:162:1–162:8, December 2009.

[2] J. Collomosse, J. E. Kyprianidis, T. Wang, and T. Isenberg. State of the 'art': A taxonomy of artistic stylization techniques for images and video. *IEEE Transactions on Visualization and Computer Graphics*, 99(PrePrints):1–1, 2012.

[3] D. DeCarlo and A. Santella. Stylization and abstraction of photographs. *ACM Trans. Graph.*, 21:769–776, July 2002.

[4] G. Di Blasi and G. Gallo. Artificial mosaics. *The Visual Computer*, 21:373–383, 2005. 10.1007/s00371-005-0292-4.

[5] R. W. Floyd and L. Steinberg. An adaptive algorithm for spatial grey scale. *Proceedings of the Society for Information Display*, 17(2):75–77, 1976.

[6] A. A. Gooch and P. Willemsen. Evaluating space perception in NPR immersive environments. In *Proceedings of the 2nd international symposium on Non-photorealistic animation and rendering*, NPAR '02, pages 105–110, New York, NY, USA, 2002. ACM.

[7] B. Gooch, E. Reinhard, and A. Gooch. Human facial illustrations: Creation and psychophysical evaluation. *ACM Trans. Graph.*, 23:27–44, January 2004.

[8] A. Hertzmann. Non-photorealistic rendering and the science of art. In *Proceedings of the 8th International Symposium on Non-Photorealistic Animation and Rendering*, NPAR '10, pages 147–157, New York, NY, USA, 2010. ACM.

[9] T. Isenberg, P. Neumann, S. Carpendale, M. C. Sousa, and J. A. Jorge. Non-photorealistic rendering in context: an observational study. In *Proceedings of the 4th international symposium on Non-photorealistic animation and rendering*, NPAR '06, pages 115–126, New York, NY, USA, 2006. ACM.

[10] H. Kang, S. Lee, and C. K. Chui. Coherent line drawing. In *Proceedings of the 5th international symposium on Non-photorealistic animation and rendering*, NPAR '07, pages 43–50, New York, NY, USA, 2007. ACM.

[11] D. Kim, M. Son, Y. Lee, H. Kang, and S. Lee. Feature-guided image stippling. *Computer Graphics Forum*, 27(4):1209–1216, 2008.

[12] H. Li and D. Mould. Contrast-aware halftoning. *Computer Graphics Forum*, 29(2):273–280, 2010.

[13] H. Li and D. Mould. Artistic tessellations by growing curves. In *Proceedings of the 6th international symposium on Non-photorealistic animation and rendering*, NPAR '11, pages 125–134, New York, NY, USA, 2011. ACM.

[14] H. Li and D. Mould. Content-sensitive screening in black and white. In *GRAPP*, pages 166–172, 2011.

[15] H. Li and D. Mould. Structure-preserving stippling by priority-based error diffusion. In *Proceedings of Graphics Interface 2011*, GI '11, pages 127–134. Canadian Human-Computer Communications Society, 2011.

[16] Y. Liu, O. Veksler, and O. Juan. Generating classic mosaics with graph cuts. *Computer Graphics Forum*, 29(8):2387–2399, 2010.

[17] R. L. Mandryk, D. Mould, and H. Li. Evaluation of emotional response to non-photorealistic images. In *NPAR*, pages 7–16, 2011.

[18] D. Mould. Stipple placement using distance in a weighted graph. In *Proceedings of Computational Aesthetics*, pages 45–52, 2007.

[19] D. Mould. Texture-preserving abstraction. In *Expressive*, pages 75–82, 2012.

[20] D. Mould, R. L. Mandryk, and H. Li. Emotional response and visual attention to non-photorealistic images. *Computers & Graphics*, 36(6):658–672, 2012.

[21] V. Ostromoukhov. A simple and efficient error-diffusion algorithm. *ACM Transactions on Graphics (SIGGRAPH 2001 issue)*, pages 567:1–572:8, 2001.

[22] W.-M. Pang, Y. Qu, T.-T. Wong, D. Cohen-Or, and P.-A. Heng. Structure-aware halftoning. *ACM Transactions on Graphics (SIGGRAPH 2008 issue)*, 27(3):89:1–89:8, 2008.

[23] Y. Qu, W.-M. Pang, T.-T. Wong, and P.-A. Heng. Richness-preserving manga screening. *ACM Transactions on Graphics (SIGGRAPH Asia 2008 issue)*, 27(5):155:1–155:8, December 2008.

[24] P. Rosin and J. Collomosse. *Image and Video-Based Artistic Stylisation*. Springer, 2013.

[25] A. Santella and D. DeCarlo. Visual interest and npr: an evaluation and manifesto. In *Proceedings of the 3rd international symposium on Non-photorealistic animation and rendering*, NPAR '04, pages 71–150, New York, NY, USA, 2004. ACM.

[26] J. Schumann, T. Strothotte, S. Laser, and A. Raab. Assessing the effect of non-photorealistic rendered images in cad. In *Proceedings of the SIGCHI conference on Human factors in computing systems: common ground*, CHI '96, pages 35–41, New York, NY, USA, 1996. ACM.

[27] A. Secord. Weighted Voronoi stippling. In *Proceedings of the 2nd international symposium on Non-photorealistic animation and rendering (NPAR)*, pages 37–43, 2002.

[28] H. Seifi, S. DiPaola, and J. T. Enns. Exploring the effect of color palette in painterly rendered character sequences. In *Proceedings of the Eighth Annual Symposium on Computational Aesthetics in Graphics, Visualization, and Imaging*, CAe '12, pages 89–97, Aire-la-Ville, Switzerland, Switzerland, 2012. Eurographics Association.

[29] M. Son, Y. Lee, H. Kang, and S. Lee. Structure grid for directional stippling. *Graphical Models*, 73(3):74 – 87, 2011.

[30] M. Sonka, V. Hlavac, and R. Boyle. *Image Processing: Analysis and Machine Vision*. CL-Engineering; 2 edition, 1998.

[31] R. Ulichney. *Digital Halftoning*. MIT Press, 1987.

[32] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13(4):600–612, 2004.

[33] H. Winnemöller, S. C. Olsen, and B. Gooch. Real-time video abstraction. *ACM Trans. Graph.*, 25(3):1221–1226, July 2006.

[34] M. Zhao and S.-C. Zhu. Sisley the abstract painter. In *Proceedings of the 8th International Symposium on Non-Photorealistic Animation and Rendering*, NPAR '10, pages 99–107, NY, USA, 2010. ACM.