Evaluation of Emotional Response to Non-Photorealistic Images

Regan L. Mandryk*
University of Saskatchewan

David Mould†
Carleton University

Hua Li‡
Carleton University

Abstract

Non-photorealistic rendering (NPR) algorithms are used to produce stylized images, and have been evaluated on the aesthetic qualities of the resulting images. NPR-produced images have been used for aesthetic and practical reasons in media intended to produce an emotional reaction in a consumer (e.g., computer games, films, advertisements, and websites); however, it is not understood how the use of these algorithms affects the emotion portrayed in an image. We conducted a study of subjective emotional response to five common NPR approaches, two blurring techniques, and the original image with 42 participants, and found that the NPR algorithms dampened participants' emotional responses in terms of arousal (activation) and valence (pleasure).

CR Categories: I.3.0 [Computer Graphics]: General;

Keywords: non-photorealistic rendering, emotion, affect, arousal, valence

1 Introduction

Non-photorealistic rendering (NPR) algorithms produce images in a wide range of expressive styles, including painting, drawing, and cartoons. These NPR images have some practical advantages over photographs, including a lack of distracting or irrelevant detail (e.g., in medical or archaeological illustrations), emphasis and clarification of crucial details (e.g., in caricature, maps, and technical illustrations), and, in some cases, ease of storage and reproduction. Beyond these practical advantages, NPR images have aesthetic benefits as the stylized images have an inherent beauty, vitality, and interest as compared to photographic images. Because of the practical and aesthetic advantages of stylized images, NPR algorithms have seen increasingly widespread use in computer games (e.g., the 2008 Prince of Persia, Borderlands, Team Fortress 2), films (e.g., *A Scanner Darkly*), and advertisements (e.g., Charles Schwab Investments), making the improvement of NPR algorithms an active area of research.

Researchers investing effort in improving NPR algorithms generally evaluate their results by considering the aesthetic quality of the resulting images. However, the use of the images for emotional media such as television, film, and advertisements means that the aesthetic quality of the image is not the only consideration of success of an algorithm—we must also consider how the resulting image elicits an emotional response from the viewer. Media creators who intend to provoke an emotional response in the viewer of a stylized image need to know if the use of a particular NPR algorithm alters the viewer's emotional response; however, researchers have little understanding of how the varying algorithms affect the perceived emotional content of the resulting image.

To determine how the perceived emotional content of the image is affected by the different algorithms, we conducted a study investigating the emotional response of 42 users to five well-known NPR algorithms (Haeberli's interactive painting [Haeberli 1990]; the photo abstraction of Orzan et al. [Orzan et al. 2007]; the abstract painting of Zhao et al. [Zhao and Zhu 2010]; Secord's weighted Voronoi stippling [Secord 2002]; and the line drawing techique of Kang et al. [Kang et al. 2007]), two blurring techniques that were used to reduce the image information in a systematic manner (uniform blur and salience-based blur), and the original image. Emotional response was measured using the three-dimensional model of emotion, which consists of arousal (level of activation), valence (positive versus negative), and dominance (level of control). Participants also rated the aesthetic quality of each image and ranked the techniques in a post-experiment questionnaire.

We found that the algorithms significantly affected user ratings of both arousal and valence. In general, the NPR techniques dampened the emotional response to the images, moving participant responses toward neutral ratings. Image abstraction yielded emotional responses closest to the original image in both arousal and valence, whereas painterly styles showed the greatest flattening of emotional responses. Differences in arousal ratings between the algorithms grew larger as the images were more arousing, whereas differences in valence ratings between the algorithms were larger for low-valence and high-valence images than for neutrally-valent images. Our results also show that the differential emotional responses cannot be solely attributed to information loss as a result of the filtering algorithms, and do not depend on whether participants had previously seen the original image. Our results are of interest to NPR researchers, but are also of particular importance for artists, designers, and media creators who use the algorithms in media intended to produce an emotional response in a consumer, such as films, games, advertisements, and websites.

2 Related Work

We first give an overview of non-photorealistic rendering and some of the work that has sought to evaluate synthetic images, followed by a primer on representations and measurement schemes for emotions, and finally a brief commentary on emotional responses to NPR.

2.1 Non-photorealistic Rendering

Since the beginnings of NPR, myriad processes have been devised for synthesizing images in a wide range of different styles and traditional artistic media, including line art, mosaics, and painterly. While some methods are interactive and depend on user input to create images, others are automated and require only a scene description, either in the form of geometry or as an input image.

Non-photorealistic images have been compared with manually created artistic images since the beginning of the field, leading Strothotte and Schlechtweg [2002] to postulate the "NPR Turing Test". In this thought experiment, people are presented with im-

*e-mail: regan@cs.usask.ca
†e-mail:mould@scs.carleton.ca
‡hli1@connect.carleton.ca

ages and asked to guess whether they were created by humans or by computers. One outcome of the observational study of Isenberg et al. [2006] was the finding that, for the algorithms tested, people generally could distinguish between images created by computer and those created by hand. The NPR Turing Test is a benchmark we have not yet passed.

Non-photorealistic rendering has had different goals, including novelty, meeting the technical challenge, and comprehensibility of the resulting images. This last point is frequently used to motivate work in the area, with medical, archaeological, and technical illustration given as application domains; perceptual studies such as those of Winnemöller et al. [Winnemöller et al. 2007; Winnemöller et al. 2006] have shown improvements in subjects' ability to recognize and comprehend objects depicted in abstracted styles compared with realistic depictions. NPR methods have generally not been cast by their creators as generating artistic or emotionally laden content; a possible exception is the work of Shugrina et al. [2006], where the perceived emotional state of the viewer drives the appearance of the image. Nonetheless, researchers in NPR often seek to improve the aesthetic qualities of their synthetic images. Zhao and Zhu [2010] use aesthetic arguments (about the appeal of abstract paintings) to motivate their algorithm, and perform a user study measuring the ease with which subjects could recognize objects in their synthetic semi-abstract paintings.

Duke et al. [2003] investigated the emotional impact of a few types of images, though in an unsystematic way and without using images that were rated for affect: the coverage of emotional space was sparse and the range of styles covered was small. Our effort here explores the impact of a range of styles, on a set of images specifically chosen to provide full coverage of a 2D emotional space. We discuss possible parameterizations of emotional space next.

2.2 Affect and Emotion

We are interested in understanding a user's emotional response to various NPR algorithms, so we must first consider how emotions are described and measured. The terms affect, emotion, and mood are often used interchangeably; we use affect to describe the low-level user responses to a stimulus (e.g., palms sweating, heart racing), emotion to describe the cognitive interpretation of the low-level responses (e.g., fear, surprise), and mood to describe the longer-term state of the user as they experience emotions. Affective responses are fleeting, emotions are short-lived, and moods change slowly over time. In this paper, we will use both affect and emotion to describe participant responses to the images.

2.2.1 Representing Emotion

There have been two main approaches to describe emotions: categorical and dimensional. The categorical approach applies specific and discrete labels to various emotions through semantic labels (e.g., sadness, pride, fear) [Ekman 2005]. The dimensional approach [Russell et al. 1989; Lang 1995] proposes that emotions can be represented by two primary orthogonal axes called arousal and valence. Valence describes the pleasure (positive) or displeasure (negative) of a feeling. Arousal is related to the energy or activation of the feeling and is typically described as low (e.g., sleepiness) to high (e.g., excitement) arousal. This arousal-valence space has been described as the circumplex model of emotion [Russell et al. 1989], and has been used to describe the categorical emotion labels. For example, 'anger' would be a high-arousal, low-valence state, while 'depression' would be a low-arousal, low-valence state. One criticism of the dimensional model is that the arousal and valence axes are not completely independent [Lang 1995]. For example, an emotion that is truly displeasurable is unlikely to also be very relaxing. A third axis, dominance, has been added to describe the feeling in terms of how controlled, influenced, or submissive it is, as compared to controlling, influential, or dominant [Coan and Allen 2007]. Together, these three dimensions have been used in emotional assessment.

2.2.2 Measuring Emotion

Although there are discrete methods of measuring emotional state (e.g., semantic differential scale) [Coan and Allen 2007], we focus on the dimensional approaches using the three-axis model of emotional state.

Based on their circumplex model, Russell et al. used the arousal-valence space to create the affect grid [Russell et al. 1989]. The affect grid is a tool for quick assessment of affect in terms of arousal and valence. Participants place checkmarks in the squares of a grid in response to different stimuli. Avoiding semantic labels, the self-assessment manikin [Bradley and Lang 1994] is a 9-point pictorial scale used to self-report arousal, valence, and dominance using a series of 5 images, with blank images between. As shown in Figure 1, the SAM provides a fast, easy, and non-linguistic way of assessing emotional state along three dimensions. Although subjective approaches are most commonly used to assess user emotional state, objective methods have been developed including using facial expression analysis, physiological signal analysis, and observational analysis (see [Coan and Allen 2007]). The studies we describe in this paper use subjective self-assessment.

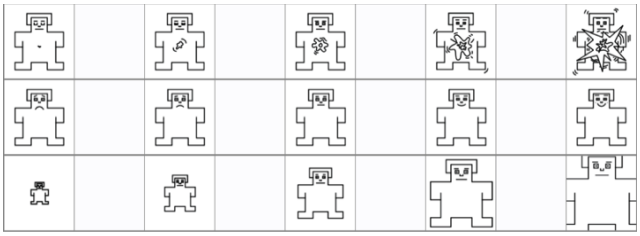

Figure 1: *The self assessment manikin 9-pt pictorial rating scale. Top: arousal; middle: valence; bottom: dominance.*

2.3 Emotional Response to Graphics

There has been little work investigating a person's emotional response to computer-generated graphics. Hertzmann [2010] motivates this issue well, pointing out that perceptual, aesthetic, and emotional content of artistic images are at least somewhat independent: "a work that is interesting is not necessarily beautiful." For 'interesting' we could substitute various other descriptive terms. There are numerous studies of the aesthetic properties of images, and perceptual studies are also common, but studies of emotional responses to NPR are rare. We previously mentioned the efforts of Duke et al. [2003] in this regard. Shugrina, Betke, and Colomosse [2006] use a 2D arousal-valence emotional space, but for image synthesis, not for evaluation of responses to other images. In fact, with their interactive system, detecting the user's emotional response to the image being created forges a probably undesirable feedback loop. Colton et al. [2008] also generate NPR images based on emotion, but use automated facial expression analysis to determine a viewer's emotional state, which in turn determines the NPR approach to use for image generation. Mar et al. [2007] investigated fundamental physiological reactions (in terms of blood oxygenation in the brain) to live-action or computer-animated agents and found that the response associated with the perception of agency was greater for the real versus the animated

agents. Although they did not control the rendering of the animated agents, they used identical film clips from a movie that had been adjusted by animators to look like cartoons.

There has been some work investigating a person's emotional response to aspects of abstract art, which may inform our understanding of emotional response to stylized images. For example, Valdez and Mehabrian [1994] describe a relationship between perceived valence and hue (wavelength), where they showed experimentally that blue was the most pleasant colour and that yellow was the least pleasant. They also showed that less bright and more saturated colours were more arousing. Their results were supported by Simmons in a study on the associations between colours and emotion [Simmons 2006]. There have also been observed relationships between shape and perceived emotion. For example, Ibanez [2011] found that perceived valence corresponded with degree of symmetry in images that had their colour held constant. Mono [1997] showed that circles, spirals, and shapes with smooth curves were more pleasant than shapes with hard angles. These results conform to earlier investigations into the emotion conveyed by drawings, where Hevner [1935] found that curves denoted serenity, while jagged strokes and harsh angles denoted fury or agitation. In a similar vein, Halper et al. [2003] found a relationship between line style and perceptions of safety—objects rendered using jagged lines were perceived as more dangerous than objects rendered using smooth lines. Line style was also related to perceptions of character strength (strong lines indicated strong characters).

3 Generating Stimuli

3.1 Choosing the Stimuli

To determine affective response to various NPR algorithms, we needed to choose a set of images to use as stimuli. The International Affective Picture System (IAPS) is a set of images that span emotional space. IAPS images were created for use as experimental stimuli, have been used in numerous studies, and provide normative ratings of emotion in terms of valence, arousal, and dominance for 956 color photographs [Lang et al. 2008].

We chose 18 IAPS images that spanned the arousal-valence space for use in our study. Images were chosen to represent 9 specific arousal-valence locations, as shown in Figure 2, including all combinations of low, neutral, and high arousal and valence. We were interested in whether the affective response to images of objects or scenes would be differentially affected by the various algorithms, so we chose an image at each of the 9 locations that represented an object (e.g., tiger, gun) and one that represented a scene (e.g., beach, cemetery) for 18 images in total.

Due to regulations of IAPS use, we are unable to publish the images used; however, Figure 2 shows a description of the image and the mean arousal, valence, and dominance of the images as provided by the IAPS documentation [Lang et al. 2008].

3.2 Rendering the Stimuli

We rendered the stimuli in several styles using existing algorithms from the literature. We selected algorithms that were capable of operating automatically on images; our source data was in the form of images, so it would not have been possible to employ approaches that require geometry, and we wanted to avoid interactive algorithms in order to avoid the possibly confounding effects of the human user working through different interfaces.

We chose to employ three main styles: stippling, line art, and painterly rendering. We also used the generic photo abstraction

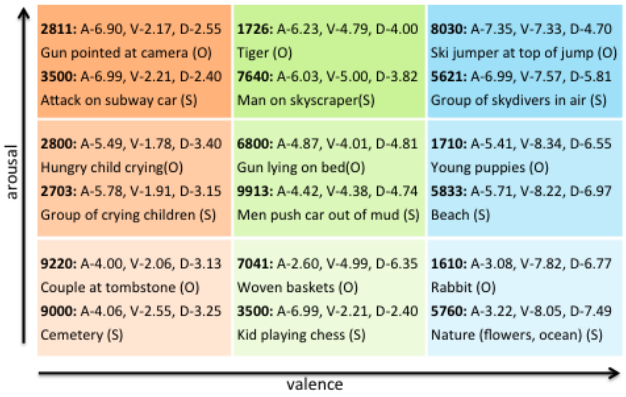

Figure 2: *The 18 IAPS image stimuli at 9 different emotional locations in arousal-valence space. The vertical axis shows increasing arousal; the horizontal axis shows increasing valence.*

method of Orzan et al., and we included two variants of blurring (uniform and salience-dependent) to provide a baseline abstraction. Stippling was one of the selected styles because of its long-standing interest to computer graphics practitioners. Line drawing and painterly rendering were chosen because of their long history and widespread use in non-photorealistic rendering. To obtain the stippled images, we used the classic stippling method of Secord [2002]. The method of Kang et al. [2007] was used to produce line art: among automatic image-based methods, the results of this algorithm are unsurpassed. For the painterly algorithms, we faced a difficult decision because of the wide variety of methods and the different styles of output they generate; ultimately we decided to use two algorithms, the classic Haeberli "Paint by Numbers" [1990] and the recent "Sisley" [Zhao and Zhu 2010]. We consider Haeberli's method to produce more representational images (closer to the original) while the output from Sisley is more abstract. Haeberli's original system had a user guiding brushstroke placement, but we used a custom implementation of this semi-automatic system in which paint strokes were chosen from a database of six possible strokes and alpha blended onto an initially black canvas at random positions. For the blurred images, we used a Gaussian filter; the uniform blur used a radius-12 filter everywhere, while the salience-adaptive blur used radius-4 filter in the background followed by a second radius-4 filtering pass everywhere. This process ensured a clearer image in the foreground but avoided the appearance of a visible boundary between more and less strongly blurred regions.

For the Haeberli, Sisley, and blurring algorithms, we used manually painted binary salience mattes to distinguish between regions of high importance and regions of low importance. The mattes were created manually; our process was for some of the authors to paint candidate mattes for the images, then review and discuss the decisions for the mattes, and finally for one person to create a final matte for each image informed by the previous discussion. Note that the same matte was then used for all three styles.

Example images illustrating the selected techniques are shown in Figures 3 and 4. Figure 3 shows the input to the algorithms: the original images and the hand-drawn salience masks. Note that these are illustrative only—the terms of use of the IAPS images do not allow them to be reproduced here. The results of the different rendering algorithms are shown in Figure 4.

Figure 3: *Input to the rendering algorithms. Above: original images. Below: hand-drawn salience masks.*

4 Experiment

4.1 Task

The task consisted of participants rating their affective response to the images, which were rendered using the various algorithms. Participants began by completing an informed consent, followed by the Ishihara Color Plate Test to screen out participants who showed colour vision deficiencies. Participants then completed a training task of two neutral images (IAPS 1602, 2530) presented on a grey background. Images were presented for 10 seconds and users were asked to describe the image verbally to the experimenter during this time. The description phase was followed by the rating phase; the image remained on the screen but was accompanied by the rating scales. After rating the image on all four scales, the user was presented with a grey mask and a submit button with instructions to press the button when they were ready to move on to the next trial.

Affective ratings were conducted using the self-assessment manikin 9-point pictorial scales [Bradley and Lang 1994]. Participants first rated the arousal of the image, followed by the valence, the dominance, and the aesthetic quality. Aesthetic quality was rated using a 9-point Likert scale. Only one rating scale appeared on the screen at any time and participants were required to provide a rating before moving on to the next rating scale (see Figure 5).

After the training task, participants were presented with the experimental system, which was identical in appearance and function to the training system and differed only in the presented images. The 18 images described in the section on choosing the stimuli were presented in a block for each of the 7 rendering techniques and the control condition (original image). The order of presentation of the 7 techniques was counterbalanced using a Latin Square to avoid any effects of order of presentation. In addition, half of the participants saw the control image prior to any of the techniques, while the other half saw the control images last. We varied the position of the control images to see whether knowing the un-retouched content of the image affected user response to the various techniques.

The 18 images were presented in the same order for each technique so that the emotional content of the images did not vary too greatly from one trial to the next. We started each block with a neutral image, and ended with a relaxing image to ease participants into a new technique and leave the block with a relaxed and positive image. Beginning in the middle of arousal-valence space, participants progressed through the images in a counterclockwise spiral

Figure 4: *Output from the rendering algorithms. From the top: stippling (Secord); line drawing (Kang); painting (Haeberli); painting (Sisley); photo abstraction (Orzan); object blur; uniform blur.*

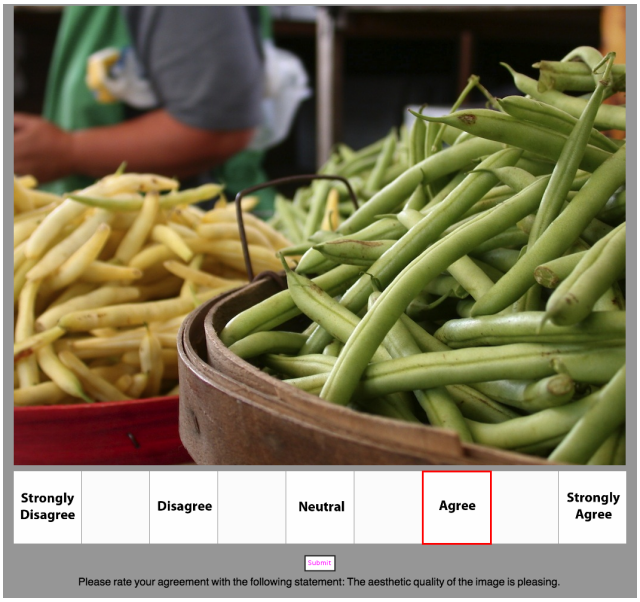

Figure 5: *Screenshot of experimental system with sample image. Participants rated the measures on a 9-pt scale by selecting the appropriate image and pressing submit.*

(see Figure 2). Since images were repeated for each of the techniques, participants were required to describe the content of the image only for the technique that was presented first.

After rating all images, participants completed a post-experiment questionnaire that gathered demographic information as well as preferences about the various techniques. The entire experiment took between 1 and 1.5 hours to complete and users were given $15 to thank them for their participation. The experiment protocol was approved by the behavioural research ethics board at the University of Saskatchewan.

4.2 Apparatus

The experiment was conducted on a Windows 7 computer and a 24 TFT display running at a resolution of 1920 by 1200. The experimental software was written in Processing. All images were presented at a resolution of 1024 by 768 pixels. The system logged the information about participants, images, and ratings in a text file for subsequent analysis.

4.3 Participants

There were 42 participants, aged 18 to 33 (mean 24), of which 21 were female. Participants all had normal or corrected-to-normal vision and did not have any colour vision deficiencies.

4.4 Data Analyses

We conducted four separate Analysis of Variance tests (ANOVAs). After aggregating the ratings for all images, we conducted an overall repeated-measures MANOVA with 1 within-subjects factors (algorithm), 1 between-subjects factor (original image seen first or last) and 4 dependent measures (arousalRating, valenceRating, dominanceRating, aestheticRating). We also coded the images into three levels of arousal (low, neutral, and high) based on their IAPS ratings, and aggregated over these three levels rather than over all images. A RM-ANOVA with algorithm (8 levels) and imageArousal (3 levels) as within-subjects factors and original image position (2 levels) as a between subjects-factor on arousalRatings will be referred to as the ArousalANOVA. A similar process was undertaken for grouping the images according to valence and conducting a RM-ANOVA on valenceRatings (ValenceANOVA). Finally, we also aggregated separately over whether the images were objects or scenes and conducted a RM-MANOVA (ObjectMANOVA) with 2 within-subjects algorithms (algorithm, object versus scene), 1 between-subjects factor (original image position) and our 4 dependent measures. For all statistical tests, when the assumption of sphericity was violated, we used the Huynh-Feldt method of adjusting the degrees of freedom. Pairwise comparisons of significant results used the Bonferonni method of correcting for multiple tests with $\alpha = 0.05$.

5 Results

In this section, we describe the results of our statistical tests. We summarize these results at the end of the section.

Do the images create the expected affective responses? To determine whether participants were responding to the affective manipulation in a predictable manner, we looked at results from our ArousalANOVA and ValenceANOVAs. The ArousalANOVA showed a main effect of imageArousal (F2,70.6=40.5, $p \approx .000$). Bonferroni post hoc comparisons revealed that each arousal grouping was significantly different (all $p < .001$). The ValenceANOVA showed a main effect of imageValence (F2,52.1=158.0, $p \approx .000$). Bonferroni post hoc comparisons revealed that each valence grouping was significantly different (all $p < .001$). Thus, the images were producing consistent and predictable affective responses in the participants (see Figure 6).

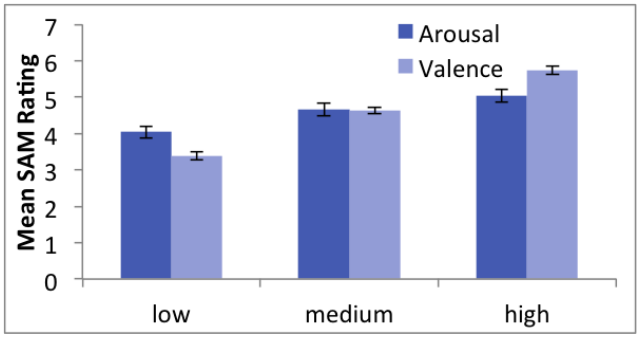

Figure 6: *Means +/- SE for arousal and valence ratings by the arousal or valence of the image stimuli.*

Are there overall differences in the affective responses to the various NPR algorithms? Using our overall RM-MANOVA, we found a significant effect of algorithm on arousalRating (F4.8,194.7=12.3, $p \approx .000$), valenceRating (F7,280=6.2, $p \approx .000$) and dominanceRating (F4.1,166.6=2.6, $p = .036$). For arousalRating, the original image was rated as more arousing than images produced with all other algorithms except Orzan. Also, the Orzan images were rated as more arousing than images produced using the Haeberli, Secord, and Sisley algorithms and images produced using the blur and object blur algorithms (see Figure 7). For valenceRating, the images produced using the Kang and Secord algorithms were rated with lower valence (more negatively) than the original images and those produced using the Orzan or Sisley algorithms. In addition, images produced using blur were rated less valent than those produced using Sisley (see Figure 7). For dominance, the images

produced using blur were rated as less dominant than those produced using Orzan or Kang (see Figure 7).

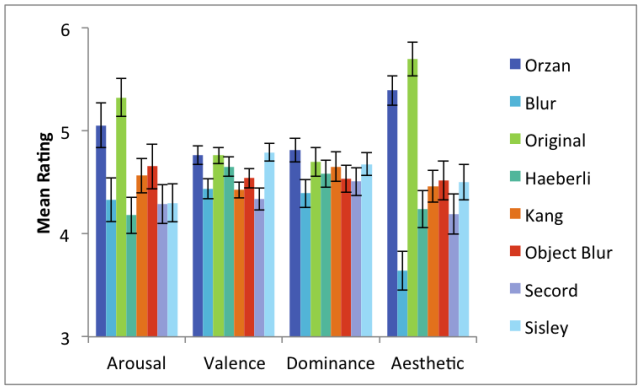

Figure 7: *Overall means +/- SE for arousal, valence, dominance, and aesthetic quality ratings.*

Were these differences affected by whether the participants saw the original image condition first or last? Our RM-MANOVA showed that there were no main effects of whether the original image was seen first or last or interactions of original image order and algorithm on any of the three affective measures (all $p > .05$).

Did the affective ratings for the different algorithms change depending on the arousal level of the original image? In addition to showing that the imageArousal was yielding consistent arousalRatings, our ArousalANOVA also showed that there was a significant interaction between imageArousal and algorithm on the arousal ratings (F4.3,489.4=11.8, $p \approx .000$). As Figure 8 shows, for low-arousal images, there was not a large difference in the arousal ratings for the different algorithms. For neutral-arousal images, the differences became larger, and for high-arousal images, the arousalRating differences were largest. Specifically, for low-arousal images, there were no significant differences between the algorithms. For neutral-arousal images, the original images were more arousing than images produced using all other algorithms except Orzan and Orzan images were more arousing than Haeberli, Sissley and blur images. For high-arousal images, the original images were more arousing than all other algorithms except Orzan, the Orzan images were more arousing than all other algorithms (except original), the Haeberli algorithm was less arousing than all except blur, and objectBlurred images were more arousing than Sisley and blur.

Did the affective ratings for the different algorithms change depending on the valence of the original image? Our ValenceANOVA showed that there was a significant interaction between imageValence and algorithm on the valence ratings (F11.0, 440.1=31.1, $p \approx .000$). As Figure 9 shows, for neutral-valence images, the differences between algorithms were small, but for low- and high-valence images, the differences between the algorithms was bigger. Specifically, for low-valence images, the Haeberli and images were rated as more valent than all others except blur, Sisley images were rated as more valent than all other algorithms except blur, and the original images were rated as less valent than all other algorithms except Orzan. In addition, blur was more valent than object blur. For medium-valence images, the only differences were that blur was rated as less valent than Orzan or Secord. For high-valence images, the original image was rated as more valent than all other algorithms, Orzan images were more valent than all but the original image, object blur was more valent than all remaining algorithms except Sisley, and Secord was rated less valent than all other algorithms except blur and Haeberli.

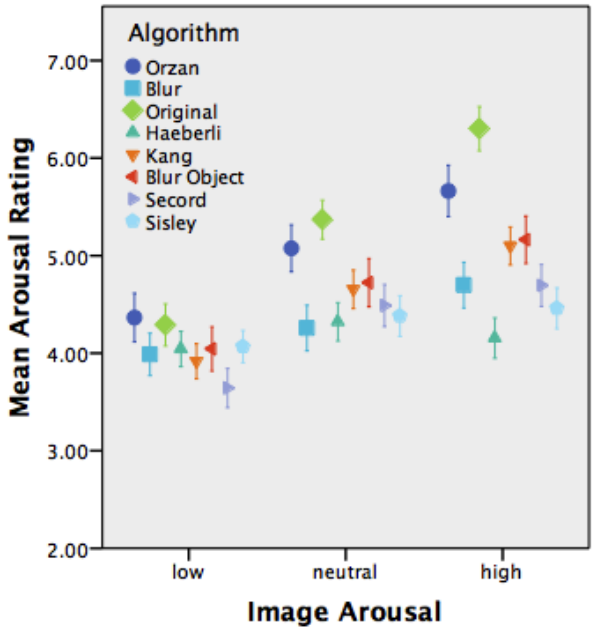

Figure 8: *Means +/- SE for arousal ratings by the arousal of the stimulus image separated by algorithm.*

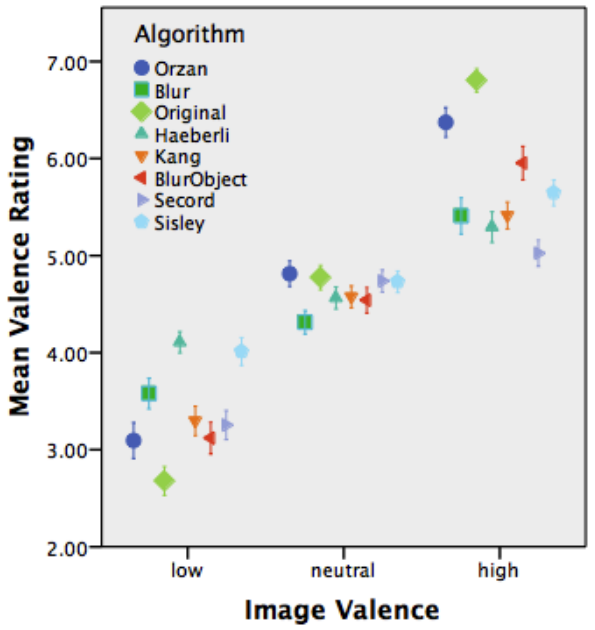

Figure 9: *Means +/- SE for valence ratings by the valence of the stimulus image separated by algorithm.*

How did the algorithms fare in terms of aesthetic ratings? The RM-MANOVA described previously showed a main effect of algorithm on aestheticRating (F5.3,211.5=23.8, $p \approx .000$). The original images and the Orzan images were rated as having a higher aesthetic quality than images produced using all other algorithms. In addition, the blurred images were rated with lower aesthetic quality than all other images except those using the Secord algorithm (Haeberli was marginal at p=.058).

Unlike the affective measures, the aesthetic ratings did change depending on whether participants saw the original images first or last. There was a main effect of original image order on aesthetic rating (F1,40=7.2, $p = .010$). Participants who saw the original images first tended to rate the aesthetic quality of all images lower on average (mean=4.3, SE=.17) than participants who saw the original image last (mean=4.9, SE=.17). This main effect needs to be interpreted in light of a significant interaction of whether the original image was seen first or last and algorithm on aestheticRating (F5.3,211.5=3.2, $p = .011$). This interaction shows that although the participants who saw the control image first rated the images as having lower aesthetic quality for all algorithms, this difference was only significant for the Orzan, objectBlur, and original image.

In addition, we also asked participants in a post-experiment questionnaire to choose their favourite and least-favorite algorithm. Participants overwhelmingly preferred the Orzan images (24/42 responses) with Sisley coming in a distant second choice (7/42). For least favourite, participants chose the blurred images most often (26/42 responses) with Secord coming in second (5/42).

Is there an overall difference in the affective ratings for images of objects and images of scenes? Our ObjectMANOVA revealed that there was a main effect of object or scene on valence ratings (F1,40=16.4, $p \approx .000$). In general, participants rated the valence of objects lower (mean=4.5, SE=.03) than the valence of scenes (mean=4.7, SE=.03). This is not surprising as the IAPS-provided valence of the object images (mean=4.8, SE=.04) was slightly lower than that of the scene images (mean=5.0, SE=.04). There were no main effects of object or scene on any of the other measures (all $p > .05$). There was also an interaction effect of object or scene and algorithm on valence ratings (F6.1, 242.6=16.4, $p \approx .000$), meaning that the ratings differences between the object and scene images were not consistent across all algorithms. Post hoc comparisons showed that there was a significant increase in the valence ratings of scenes over objects for all algorithms (all $p < .005$) except for Kang and Secord, where the valence ratings for scenes were lower (see Figure 10). There were no interaction effects of object or scene and algorithm on arousal ratings or dominance ratings (all $p > .05$).

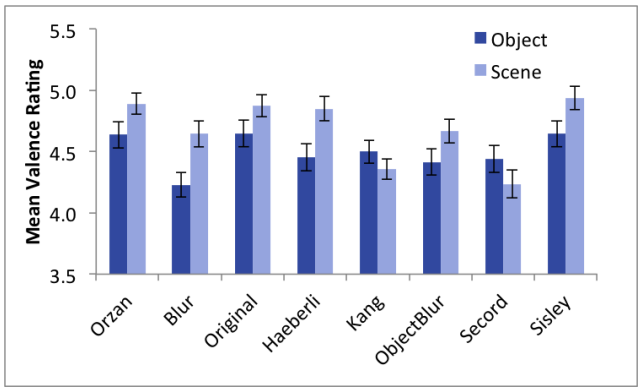

Figure 10: *Means +/- SE for valence ratings by algorithm and whether the stimulus image was an object or a scene.*

Summary of the results Our results can be summarized into the following nine takeaway messages:

1) The image choices were producing significant and predictable differences in the affective ratings, showing that our experimental stimuli were effective;

2) Applying any of the algorithms except Orzan created less arousing images than the original, and the Orzan algorithm created more arousing images than all of the algorithms except Kang;

3) The differences in arousal ratings between the algorithms became more apparent as the image itself was more arousing. There were no differences in the algorithms for low-arousal images, small differences for medium-arousal images, and large differences for high-arousal images;

4) Images produced using Kang and Secord were less valent than images produced using Orzan, Sisley, or the original image;

5) The differences in valence ratings between the algorithms were larger for low-valence and high-valence images than neutrally-valent images;

6) For Kang and Secord, the valence ratings were higher for objects over scenes, whereas the opposite was true for all other algorithms;

7) The difference in arousal and valence ratings did not depend on whether participants saw the control images first or last;

8) Participants preferred the Orzan and original images over all other algorithms and least liked the blurred images; and

9) The differences observed in all results cannot be solely attributed to information loss, as shown by our blurred image and blurred object algorithms.

6 Discussion

Our most general finding was that the rendered images produced flattened affect, as compared with the original images: arousal was reduced and valence was brought closer to neutral. This result was consistent across all the algorithms we tested, although some algorithms more strongly exhibited this flattening pattern. It is tempting to attribute this outcome to a failure on the part of the participants to recognize the content of the stylized images due to information loss; however, there are two reasons why information loss does not fully explain our results. First, were this explanation correct, we would have observed an effect of order of presentation of the original image on the affective ratings—participants who saw the original image first would have exhibited an improved ability to interpret stylized images. The absence of an order effect for any of our affective measures is evidence against this explanation. Second, we included the blurred images and blurred background images to specifically test the possibility that our results could be attributed to information loss. There was no consistent pattern where responses to the stylized images followed responses to the blurred images. In fact, affective responses to some algorithms more closely mirrored responses to the original image (e.g., Orzan). It is possible that some of the observed affective dampening can be attributed to information loss in the stylized images; however, this is not the sole explanation and there are other factors that must be considered.

It might not be too surprising that synthetic non-photorealistic images do not have much emotional content. NPR has long been viewed as a scientific endeavour and technical challenge, and researchers have not often explicitly sought to induce emotional responses with their images. Nonetheless, work in the field has occasionally been motivated by the idea of creating

more emotionally-charged images (e.g., the recent work of Lopez-Moreno et al. [2010]). The experimental data indicate that for a broad range of existing techniques, synthetic images have less emotional impact than the photographs from which they were derived. This points to an open problem for practitioners to address.

Among the algorithms tested, the photo abstraction method of Orzan et al. produced affective responses most similar to the original. Unlike the other methods which hid or removed most detail, this method preserved details, including color gradients, region boundaries, and some high-frequency features; we speculate that the inclusion of a few details in addition to the large-scale content was responsible for this algorithm's success at evoking affective responses from the viewers. Although there is no advice in the literature on how the level of detail in images influences affective response, it stands to reason that preserving details will aid in preserving the emotional impact of the images.

There is a difference between the responses to "object" images (those concentrating on a distinct object or person) compared with the responses to "scene" images (those where large parts of the image are needed to establish context). For almost all rendering styles, and for the original images, scenes produced higher valence ratings than objects. This is not unexpected as the original images of scenes were rated with a slightly higher average valence in the IAPS database than the images of objects, and this difference may have simply carried over to the stylized images. It could also be that the loss of detail has greater impact on affective response to objects than scenes. For example, an image of a bunny might be more affected by the loss of detail in the algorithms than an image of a beach scene, where the general idea and tone of the image can be conveyed with much less detail. The differential affective response to stylized images of objects and scenes warrants future research, including questioning why this valence difference between objects and scenes was reversed in the case of the line drawing and stippling styles.

The line drawing and stippling styles are quite different: line drawing shows edges and largely preserves high-frequency details, while the stippling method indicates tone and better preserves low-frequency content. Nonetheless, both methods produced similar responses overall. We might attribute this to lack of color, that being the main commonality between the two styles.

Although we may have expected similar responses to the images that were blurred overall and those where the background was blurred more than the primary object, there is considerable difference between the uniformly blurred images and those informed by the mask. In terms of the aesthetic judgement, while the uniformly blurred style was by far the least liked, the object-blur style was competitive with the other sophisticated rendering algorithms, apart from Orzan et al.'s photo abstraction. Object blur also provided images with marginally higher arousal than most styles, as compared with uniform blur which yielded the least arousal. Altogether, this provides some support for the commonly held belief that image abstraction should pay attention to the content: less important content can be more abstracted than the more important content. Speculatively, the specific content may matter less than the simple fact of choosing some coherent subject for the image and portraying it more prominently than the background.

As a minor observation, we point out that the Sisley images did not have higher valence than the images produced by other algorithms. We had expected that the color saturation shift employed in this algorithm would have an effect: the brighter colors would have seemed more cheerful to the participants, perhaps manifesting as an increase in reported valence for the neutral images. However, no such effect was detected. Figure 7 shows that the mean valence ratings for Sisley images was comparable to Orzan and the original images; however, Figure 9 further shows how the valence benefits of Sisley images were seen mainly for the low-valence images. Like Haeberli images, Sisley images showed considerable valence dampening towards neutral ratings (i.e., higher valence ratings for low-valence images and lower valence ratings for high-valence images).

7 Conclusions

This paper investigated emotional responses to computer-generated non-photorealistic images. We conducted a 42-subject study measuring valence, arousal, dominance, and aesthetics over a set of 18 images rendered in eight different styles: five existing image-based algorithms were used, plus two variants of blurring, plus the original photographic image. The 18 original images were from the IAPS dataset and had been rated for affective content; our participants' responses were consistent with the initial rating.

We found that the use of NPR algorithms significantly affected participants' reported experiences of valence and arousal. Across all algorithms, emotional responses were muted, being shifted from more strongly felt emotions towards a neutral state. Nonetheless, the emotional responses were never suppressed entirely, nor are the reduced intensities of emotions attributable to loss of detail in the rendered images. Among the algorithms investigated, the photo abstraction of Orzan et al. best preserved emotional responses, while the painterly algorithms exhibited the most dampening. We hope that these results will provoke further investigation of emotional responses to NPR images, and that they will inspire researchers in the NPR community to devise techniques that can retain or even amplify the emotional content of the input.

7.1 Future Work

Our results are the first to show that people's emotional reactions to stylized images change with the use of different NPR algorithms. These results open a number of research opportunities in this space.

Subjective evaluation is a good approach for understanding participants' attitudes and opinions and provided significant and consistent results in our study. Prior research has shown that emotional responses to pictorial stimuli from the International Affective Picture System can also be measured via objective physiological response [Lang et al. 1993]. One main advantage of physiological measures of emotional response is that the affective response is accessed directly and not mediated by cognitive processes. We plan to add the objective measurement of emotional reaction via physiological measures to determine if there are low-level responses to the various NPR algorithms.

Our study examined reaction to still images; however, NPR algorithms have been used in emotionally-rich animated media such as computer games, films, and advertisements. We plan to extend our work to examine emotional response to animated clips of stylized images, and to consider more ecologically-valid stimuli such as dramatic or narrative film clips that include sound.

Acknowledgements

This work received financial support from NSERC and from GRAND. We would like to thank Brett Taylor, Brett Watson and all of our participants. We would also like to thank the authors of the algorithms used in the study for making their implementations available.

References

BRADLEY, M. M., AND LANG, P. J. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry 25*, 1, 49–59.

COAN, J. A., AND ALLEN, J. J. B. 2007. *Handbook of emotion elicitation and assessment*. Oxford university press.

COLTON, S., VALSTAR, M. F., AND PANTIC, M. 2008. Emotionally aware automated portrait painting. In *Proceedings of the 3rd international conference on Digital Interactive Media in Entertainment and Arts*, ACM, New York, NY, USA, DIMEA '08, 304–311.

DUKE, D., BARNARD, P., HALPER, N., AND MELLIN, M. 2003. Rendering and affect. *Computer Graphics Forum 22*, 3, 359–368.

EKMAN, P. 2005. *Basic Emotions*. John Wiley & Sons, Ltd, 45–60.

HAEBERLI, P. 1990. Paint by numbers: abstract image representations. In *Proceedings of the 17th annual conference on Computer graphics and interactive techniques*, ACM, New York, NY, USA, SIGGRAPH '90, 207–214.

HALPER, N., MELLIN, M., HERRMANN, C. S., LINNEWEBER, V., AND STROTHOTTE, T. 2003. Towards an understanding of the psychology of non-photorealistic rendering. *Comp Visualistics Media Informatics and Virtual Communities 11*, 67–78.

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

HEVNER, K. 1935. Experimental studies of the affective value of colors and lines. *Journal of Applied Psychology 19*, 4, 385 – 398.

IBANEZ, J. 2011. Showing emotions through movement and symmetry. *Computers in Human Behavior 27*, 1, 561 – 567. Current Research Topics in Cognitive Load Theory, Third International Cognitive Load Theory Conference.

ISENBERG, T., NEUMANN, P., CARPENDALE, S., SOUSÃ, M. C., AND JORGE, J. A. 2006. Non-photorealistic rendering in context: an observational study. In *Proceedings of the 4th international symposium on Non-photorealist ic animation and rendering*, ACM, New York, NY, USA, NPAR '06, 115–126.

KANG, H., LEE, S., AND CHUI, C. K. 2007. Coherent line drawing. In *Proceedings of the 5th international symposium on Non-photorealistic animation and rendering*, ACM, New York, NY, USA, NPAR '07, 43–50.

LANG, P. J., GREENWALD, M. K., BRADLEY, M. M., AND HAMM, A. O. 1993. Looking at pictures: affective, facial, visceral, and behavioral reactions. *Psychophysiology 30*, 3, 261–273.

LANG, P. J., BRADLEY, M. M., AND CUTHBERT, B. N. 2008. International affective picture system (IAPS): Affective ratings of pictures and instruction manual.

LANG, P. J. 1995. The emotion probe. studies of motivation and attention. *American Psychologist 50*, 5, 372–85.

LOPEZ-MORENO, J., JIMENEZ, J., HADAP, S., REINHARD, E., ANJYO, K., AND GUTIERREZ, D. 2010. Stylized depiction of images based on depth perception. In *Proceedings of the 8th International Symposium on Non-Photorealistic Animation and Rendering*, ACM, New York, NY, USA, NPAR '10, 109–118.

MAR, R. A., KELLEY, W. M., HEATHERTON, T. F., AND MACRAE, C. N. 2007. Detecting agency from the biological motion of veridical vs animated agents. *Social Cognitive and Affective Neuroscience 2*, 3, 199–205.

MONO, R. 1997. Design for product understanding. the aesthetics of design from a semiotic approach. 2.

ORZAN, A., BOUSSEAU, A., BARLA, P., AND THOLLOT, J. 2007. Structure-preserving manipulation of photographs. In *Proceedings of the 5th international symposium on Non-photorealistic animation and rendering*, ACM, New York, NY, USA, NPAR '07, 103–110.

RUSSELL, J. A., WEISS, A., AND MENDELSOHN, G. A. 1989. Affect grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology 57*, 3, 493–502.

SECORD, A. 2002. Weighted voronoi stippling. In *Proceedings of the 2nd international symposium on Non-photorealistic animation and rendering*, ACM, New York, NY, USA, NPAR '02, 37–43.

SHUGRINA, M., BETKE, M., AND COLLOMOSSE, J. 2006. Empathic painting: interactive stylization through observed emotional state. In *Proceedings of the 4th international symposium on Non-photorealistic animation and rendering*, ACM, New York, NY, USA, NPAR '06, 87–96.

SIMMONS, D. R. 2006. The association of colours with emotions: A systematic approach. *Journal of Vision 6*, 6, 251.

STROTHOTTE, T., AND SCHLECHTWEG, S. 2002. *Non-photorealistic computer graphics*. Morgan Kaufman Publishers.

VALDEZ, P., AND MEHRABIAN, A. 1994. Effects of color on emotions. *Journal of experimental psychology General 123*, 4, 394–409.

WINNEMÖLLER, H., OLSEN, S. C., AND GOOCH, B. 2006. Real-time video abstraction. *ACM Trans. Graph. 25* (July), 1221–1226.

WINNEMÖLLER, H., FENG, D., GOOCH, B., AND SUZUKI, S. 2007. Using NPR to evaluate perceptual shape cues in dynamic environments. In *Proceedings of the 5th international symposium on Non-photorealistic animation and rendering*, ACM, New York, NY, USA, NPAR '07, 85–92.

ZHAO, M., AND ZHU, S.-C. 2010. Sisley the abstract painter. In *Proceedings of the 8th International Symposium on Non-Photorealistic Animation and Rendering*, ACM, New York, NY, USA, NPAR '10, 99–107.