Authorial Subjective Evaluation of Non-Photorealistic Images

David Mould*
Carleton University

Abstract

I argue in favor of a systematic subjective evaluation of non-photorealistic images. Objective measurements are hard to design, and quantitative user studies are problematic for a multiplicity of reasons. Subjective evaluations are not quantitative but are faster to conduct and offer the chance to dig into subtleties that are obscured by numerical scores. By carefully laying out the important elements of the intended image style, and then evaluating their results according to their adherence to the style, researchers can produce convincing evaluations with a manageable level of effort.

CR Categories: I.3.6 [Computer Graphics]: Methodology and Techniques— [J.5]: Arts and Humanities—Fine arts

Keywords: non-photorealistic rendering, aesthetics, evaluation, computer-generated art, user studies

1 Introduction

How can we evaluate results in non-photorealistic rendering? This question is one of the grand challenges of the field [Gooch et al. 2010], as posed by Salesin in 2002; in the time since, we have made little progress in addressing it. One possibility is to deploy user studies, a trend which Hertzmann [2010] inveighs against; another is to eschew measurement in favor of appreciation, as advocated by Hall and Lehmann [2013]. In this paper, I attempt to describe a structured approach to subjective evaluation, sometimes paralleling the arguments given by Hall and Lehmann but coming to quite different conclusions.

I take for granted that we need to evaluate our results in some way. We need to be able to distinguish between worthwhile results and those that need more attention before being brought to a wider audience through publication. I hope this is uncontroversial; the question is not whether evaluation should proceed, but what form evaluation should take.

The method of evaluation needs to be tailored to the type of result [Isenberg 2013]. User studies are well suited to answering questions about quantifiable goals, as I will discuss later. Methods that are primarily tools can be evaluated according to the dictates of the task with which the tool assists. Rendering styles with quantifiable objectives – making the content of the images more recognizable, for example [Winnemöller et al. 2006] – can also be evaluated with user studies. When objective methods of evaluation are available, these should be preferred; however, objective measurements in NPR are often proxies for quantities of interest rather than being able to provide definitive answers directly.

*mould@scs.carleton.ca

This paper concentrates on the thread of NPR research which creates *interesting images*, by which I mean synthetic images created with artistic or aesthetic intent but lacking a pragmatic purpose. Such images are said by Hall and Lehmann [2013] to have the purpose of "being art", following a broad view of art such as that of McCloud [1993], for whom everything is art that is made by humans without a direct purpose. Typically, NPR in this vein creates images made to resemble artistic images within a known historical style or made using traditional media. Styles of generic abstraction also qualify, such as those made by DeCarlo and Santella [2002] or Kyprianidis [2011], as do computational approaches to relatively novel artforms such as photomosaics [Tran 1999].

In writing this paper, I have in mind methods that generate images mostly automatically, with minimal human intervention. I will focus on the evaluation of the images themselves, rather than the underlying algorithms. Evaluation of algorithms can be conducted along dimensions such as computation time, elegance, scalability, and robustness, independently of the images created. An algorithm's *correctness* is arguably its important aspect, and we would only bother to evaluate runtime and other properties once the correctness of the output has been established. In our context, evaluating the quality of the images is roughly equivalent to establishing whether or not the output should be considered correct.

In the earliest days of NPR, and even sometimes today, evaluation can be done implicitly, simply by showing results: "Behold!" This is unsatisfying, however. In the case that the reader disagrees with the implied positive judgement, there is no fallback. The researchers should strive to supply the reader with insight about why the judgement should be positive; dissecting the results and identifying good and bad aspects is helpful. More recently, one standard evaluation method is the user study, but user studies are time-consuming and, in NPR, often perfunctory, since this is not our area of expertise.

User studies function best when there is a *user* that is set an identifiable *task*. Both qualitative and quantitative approaches follow the user through the performance of the task: quantitative approaches measure something about the effectiveness of the method (e.g., completion time, error rate) while qualitative approaches ask the user or an observer to respond to questions in the context of the task. Images do not have a user: they have an audience. Hence, task-oriented evaluation methods are not suitable.

I advocate a structured qualitative analysis of images, where goals are clearly articulated and then results are described with specific reference to the intial objectives. As mentioned above, the images will be generated without any particular task in mind, but rather a general objective of provoking the interest of the audience. Articulating more specific goals forces the researchers to identify the key characteristics of the target image style; the researchers and the readers can then look for these characteristics in the final images. Creating the list of characteristics might be thought of as building a theory of the style. After constructing the algorithm and producing some output images, the researchers can search for the listed characteristics in a systematic way. Holistic evaluation will also be necessary, but the list provides some structure to the discussion.

Crucially, this style of evaluation mandates a high level of disclosure. The researchers' initial goals must be laid out explicitly; the target style should be described in detail and justified for the reader.

If the algorithm omits a certain feature visible in the targeted style, the omission should be discussed. The reader can follow the reasoning and agree or disagree, ideally being able to identify specific points of disagreement. Such evaluations will necessarily be subjective, but more substantial than vacuous statements equivalent to "we liked it".

One thread of conventional wisdom has been that user studies can substitute for subjective evaluation. Of course, the users' reactions themselves are subjective, but (the hope is) unbiased: they will not be inclined to favor the authors' results unduly. Thus, by averaging out many subjective opinions, we obtain objectivity. Unfortunately, the objectivity thus obtained is illusory. Being able to attach numbers to the responses does not make the numbers meaningful, nor the results of arithmetical operations on the numbers. The users may be uniformly biased, and indeed are likely to be unless specific steps are taken to forestall the bias. In many cases, we are asking the subjects to make careful, subtle judgements: ones they have neither the capacity nor the inclination to do properly. I discuss user studies at more length later in this paper.

One of the chief virtues of applying a systematic subjective evaluation is that it allows later readers to verify the results: everything is presented, and the reader can inspect the list of criteria for omissions, read the justifications to see why certain possibly-doubtful items were included, and inspect the final images to verify that the initial promises were met. The reasoning is laid out, and the reader can follow along and confirm that the argument makes sense – or not. Exposing the reasoning to scrutiny allows later researchers to spot the gaps and make new methods to fill them, and to compare results in a nuanced way that quantitative methods generally do not allow.

1.1 Beyond Appreciating

Hall and Lehmann [2013] make the case that NPR should borrow evaluation techniques from art history, somewhat echoing remarks made by Greenberg and Buxton on evaluation in HCI. They exhort "Don't measure – appreciate!". I agree with the first part. I wonder, though, whether their intent would be better captured by asking for art *criticism* rather than art *appreciation*. Criticism, in the sense of film criticism or music criticism, refers to reporting an evaluation and judgement and the reasoning supporting them. Kosara et al. [2008] advocate a critical approach to visualization in a similar register, albeit primarily as an element of an iterated design strategy as opposed to an evaluation method to be applied to novel visualizations.

Appreciation, unlike criticism, is both passive and instinctive; Pepper [1949] remarks that "to appreciate [something] is to find delight in it... [as] a 'thing of beauty'." Pepper opposes the appreciative mode to the practical mode, where things are valued for their uses, and to the analytical mode, where things are studied and classified. Criticism partakes of the analytical mode, which arrives at conclusions and recommendations through reasoning. Appreciation can fuel criticism, but must first be investigated and explained: it is not sufficient by itself.

One concern of Hall and Lehmann is the relationship of NPR to fine art. They state that "a commentary [on social issues] is essential for art appreciation", privileging the perspective that Kushner [1983] terms the "message approach" to art. In contrast, the "structural approach" [Kushner 1983] makes judgements strictly according to the intricacy and excellence of the forms in the art, and does not attempt to grapple with the artist's intent or message. Social commentary in art is a thorny issue and largely beyond the scope of this paper, but I did want to note the existence of a school of thought that holds that images can be evaluated on the merits of their internal structure, without reference to the symbols and concepts behind the image. NPR traditionally has operated with a structural approach to both image synthesis and evaluation.

Hall and Lehmann point out that NPR practitioners sometimes fail to characterize their objectives adequately; phrases such as 'the style of the Impressionists' come under criticism on the grounds that there is no such singular style. Hall and Lehmann contend that NPR practitioners should be trained in art history, since such training would make evaluation easier and would prevent blunders such as the quoted phrase. Desirable as it might be, I am not optimistic that this vision will be realized in the near future; in the remainder of this paper, I offer a naive systematic approach to evaluation that does not depend on any specific knowledge of art history, but rather an eye for detail and a willingness to be thorough. These traits should be possessed to some degree by all computer graphics practitioners.

1.2 On Evaluation in Visualization

Researchers in the field of data visualization have faced a similar problem to ours: it is hard to evaluate a novel visualization technique. Further, information visualizations are often designed to have aesthetic appeal as well as meeting practical goals. The existence of practical goals does set apart information visualization applications from artistic renderings, though. An information visualization has the objective of communicating features and relationships in the data. A visualization should also not mislead a viewer into perceiving a relationship that is not in fact present.

Because the practical objectives have measurable outcomes, quantitative user studies can be an appropriate mechanism for evaluating visualizations. Nonetheless, as reported by Isenberg et al. [2013], quantitative studies are done in a minority of papers; by far the most common evaluation technique is 'qualitative results inspection', QRI, an informal approach where sample results are shown and the reader invited to marvel at them. QRI is roughly equivalent to the evaluation method called *visual inspection* in this paper.

Isenberg et al. lament the visualization field's apparent emphasis on quantitative studies. They suggest that qualitative studies, such as structured interviews with experts, are more suitable for the often broad and ill-defined visualization tasks that their field attempts to address. Their emphasis on rigor in qualitative assessment is very much in line with my recommendations here.

2 Principled Subjective Evaluation

Researchers in NPR have used several different evaluation methods with a wide range of sophistication. Direct presentation of results and implicit evaluation by readers is the least sophisticated method in widespread use in graphics. In photorealistic computer graphics, comparisons can be made to ground truth, but this is rarely practical in NPR. More sophisticated evaluation methods include user studies and statistical measurements of image properties. User studies are time-consuming, especially when conducted well, and when a study is done carelessly it is easy to misinterpret the results. Direct objective measurement of properties of the images is hard to apply generally, since it necessitates a customized metric function estimating the quantity of interest. Here, I argue for a systematic approach to subjective evaluation. The key to making this idea work is to be specific and thorough in initially describing objectives, and then to return to the same list of objectives in the evaluation.

My suggested structured evaluation process consists of the following four stages:

- compile a list of the characteristics of interest

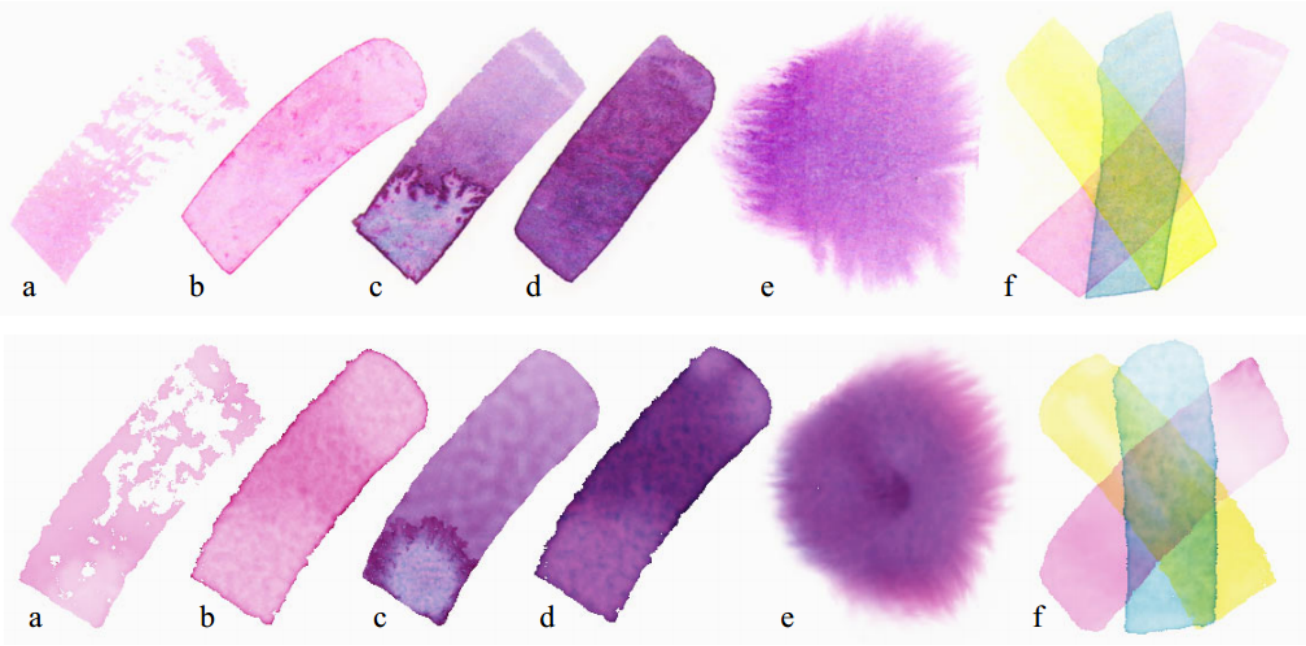


Figure 1: *Curtis's [1997] summary of simulated phenomena. Above: real watercolor; below: synthetic images. Copyright 1997 ACM, Inc. Used by permission.*

- identify the most important features and justify your assessment
- show varied examples of your results
- evaluate the examples according to their adherence to the important features mentioned

I discuss each of these points in turn.

Compile a list of the characteristics of interest. Break down the target style into components that can, in principle, be separately evaluated. This characterization of the style will drive algorithm design. Later, the output images will be evaluated according to their adherence to the items on the list.

Identify the most important elements on the list and justify the assessment. Some elements may be deemed more central to the style than others. Pragmatically, some elements may be omitted because they are too difficult to do automatically – e.g., they demand semantic understanding. The outcome of this stage is a list of final objectives for the method.

These first two stages are intended to be done prior to creating the method; indeed, by following with a stage along the lines of "create an algorithm that respects these important features", the process could serve as a manifesto for a certain style of NPR research. We are primarily concerned here with the evaluation of the research, but of course research is often an iterative process, where progress stumbles and retreats and then seeks a better path. Evaluation of early results is necessary in order to judge the promise of a research direction; that evaluation needs to be as light as possible so as not to drag down the rate of progress.

Just as the approach to the problem can change, the researchers can reformulate the problem or reorient towards a different problem in response to observations. Indeed, some of the most exciting times in research come when the work is discovered to have unlooked-for relevance to another problem entirely. If we have the ill-formed goal of creating "interesting images", we may often encounter relevant results – unintentionally interesting in an unexpected way. Such results are often spurious or non-replicable, but sometimes indicate a possible discovery.

Should a serendipitous discovery manifest, researchers must not feel beholden to the original problem. Instead, they can use the discovery as inspiration for a new research agenda, complete with a new problem statement – one to which a solution has already been found. However, the researchers must take care in formulating the new problem as carefully as the old. This not being done, sometimes work can be dismissed as "a solution in search of a problem": the inadequately reformulated problem is uninteresting, or, worse, is a problem that would be better approached using previously existing methods. When adapting a problem statement to match promising partial results, it is vital to think critically about possible alternative approaches.

Having created a method, possibly iteratively refining the method and objectives, the method should be used to create output images. The images will be evaluated visually by the researchers, who should strive for specificity and thoroughness in their comments and impartiality in their judgements. The remaining two stages of the process relate to conducting this evaluation.

Show a range of examples. In this phase, the researchers will prepare several example images illustrating different aspects of the work, using different compositions, different types and densities of features, perhaps different palettes and different ranges of tone and contrast. These examples will be examined by the readers as visual evidence supporting the paper's claims. They serve as the raw material for the systematic evaluation in the final stage.

Discuss the examples with respect to the previously identified elements of interest. The researchers should draw attention to details in the examples that showcase the method's success. These should be as specific as possible. The researchers are amassing evidence for and against the proposition that the algorithm succeeded. The

reader will form an independent opinion, guided by the authors' discussion. Likely the reader will examine the images less thoroughly than the researchers have; it is therefore incumbent on the researchers to point out the telling details that might be overlooked in a casual inspection.

Likely most if not all researchers in NPR already undertake a process akin to that described. I am in part arguing for more thorough disclosure of the process. By making assumptions and objectives explicit, researchers help the reader to evaluate the research. Are the assumptions plausible? Are the stated objectives worthwhile? Is the list of characteristics complete? When following this process, the researchers' beliefs are written out, and the reader can assess them and agree or disagree. Further, because the beliefs are written down in detail, the reader can narrow down the point of disagreement, if any. In a paper-reviewing scenario, for example, this can help the authors and referees reach consensus on what needs to be done to get the paper ready for publication. When possible, the researchers should attribute image details to elements of the algorithm. The algorithm is likely not a monolith: elements can be separated out to be reused elsewhere by others. Conversely, the authors should disclose gaps and flaws, which may represent opportunities for future work.

2.1 *Computer-generated Watercolor*

Curtis et al. [1997] provide a template for how to do the first part of the subjective evaluation. They examined real watercolors and identified several phenomena; the visual summary they created is reproduced in Figure 1. This summary provided a clear statement about the effects they intended to treat in their own work. A series of synthetic images parallel to the real ones provided the basis for evaluation using visual comparison.

Curtis et al. provided the first algorithm for synthetic watercolor, so were able to demonstrate an advance with minimal evaluation. Their actual evaluation is implicit: they invite the reader to compare the list of phenomena in the figure, with the expectation of a favorable judgement. Even a weak approximation of the phenomena would have been progress: as pioneers, Curtis et al. were in a position to define the state of the art. Their results were excellent, standing up very well today. Nonetheless, viewing these results with a dispassionate eye reveals some shortcomings. For example, the synthesized flow is too regular (part c) and the Perlin noise too prominent (parts c, d). A modern paper on watercolor rendering would be expected to compare with previous work and to render judgement, either demonstrating improved quality or acknowledging a lack of improvement but making a case for the usefulness of the proposed algorithm on other grounds, such as improved rendering speed.

I by no means intend criticism of *Computer-generated Watercolor*. On the contrary, it is because of the high bar set by early works such as this that we now have such a difficult time making the case that new results have superior quality. Curtis et al. did not dwell on the quality of their results; to some extent, it was unnecessary for them to do so, because their results did speak for themselves. Explicit discussion was also less crucial because the results were organized in such a way as to make the elements extraordinarily clear. Future researchers can emulate this clear breakdown of results while going beyond the implied evaluation to make an explicit argument about the good and bad aspects of their results as they see them. Indeed, this is sometimes done in modern papers – but not always. I would like to see specificity in qualitative judgement become universal rather than exceptional.

3 Alternatives

Broadly speaking, results in NPR are evaluated in one of three ways: first, visual inspection, possibly aided by side-by-side comparisons with similar results; second, automated measurements; third, measurements derived from user studies. I discuss each of these in turn.

Visual inspection was the primary evaluation approach in the early days of the field, as noted by Hertzmann [2010]. Hertzmann suggests that visual inspection becomes more difficult as algorithms improve and the differences between images become subtler, impelling researchers to seek objective validation. No doubt the task of persuading readers of improvements becomes more difficult as differences shrink and more nuanced judgements become necessary. Yet it is precisely under these circumstances that expert human judgement beomes most valuable. The principled subjective evaluation outlined in this paper is an effort to make human judgement more systematic and thorough.

Side-by-side comparisons to previous work are vital where previous work exists, which is increasingly common as NPR accumulates history. They are a necessary component of visual inspection. Ideally, the authors would point out specific differences between images produced by different algorithms. To the extent that side-by-side comparisons can be considered an experiment, the experiment should be controlled: the only differences should be due to differences in the algorithms. For example, the input data should be the same. This is not always possible, but for certain classes of method – those based on image processing, for example – using exactly the same input is feasible.

Image comparisons can be made not only between new results and previous work in NPR, though, but also between new results and the artistic images that inspired the work. It is common to include sample artistic images in a paper's introduction, as an aid to explaining the objectives, but closing the loop by returning to the artistic image in the evaluation is not always done.

Automated measurements come in many types. When available and relevant, they are extremely useful; for some image processing applications such as noise removal, where ground truth is available, they are ideal. Unfortunately, ground truth of image aesthetics is not normally available. Identifying a quantifiable objective function for a given application is usually difficult enough that when it can be done, we can automate an optimization process and have a contribution. The halftoning process of Pang et al. [2008] is an example. Szeliski et al. [2006], writing about Markov random fields in computer vision, advise their readers to search for new energy functions to optimize: optimization techniques are already powerful enough that the quality of the results, compared to ground truth, owes much more to differences in the energy function than differences arising from the optimization process.

Conversely, measurements of doubtful relevance can be useful in partially validating our methods and in inspiring algorithm creation, but are insufficient by themselves. If the results of Lee et al. [2006] did not resemble Pollock paintings in visual inspection, we would not accept the fractal dimension measurement as definitive. Hertzmann [2010] coined the term "proxy metrics" to describe measurements that relate to features of interest but do not capture the whole story. Implicitly or explicitly, the researchers must connect the proxy metric to the quantity of interest. Failure to do so brings to mind the proverbial drunk searching for his keys under a streetlight, far from where he lost them: "Why are you looking here, then?" – "It's too dark over there."

Even when the proxy is initially connected to the actual goal, attempting to optimize for the proxy can lead it away from the actual

goal. This is akin to Goodhart's Law in economics, where a statistically observed correlation can be destroyed by efforts to manipulate one of the correlates. In the case of NPR research, an unreflective reliance on automated measurements can lead the search away from our actual goals: it can inhibit investigation of questions for which automated measurements will be difficult. At the same time, it can encourage researchers to investigate questions where the results are easy to measure, regardless of the importance of these questions.

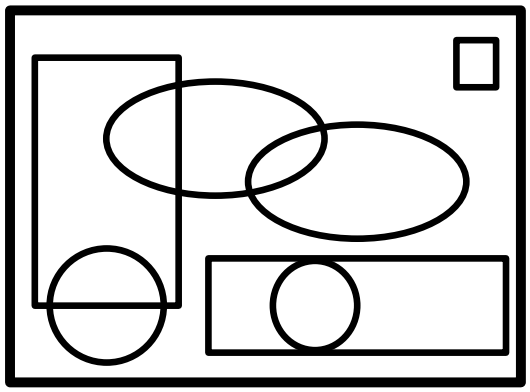
User studies are a quagmire that I discuss at greater length in the following section. User studies are appealing because they seem to offer the best of both worlds: nuanced human judgement freed from subjectivity by force of numbers. As Hertzmann previously argued, and I argue below, this apparent objectivity is illusory. Where user studies generate proxy metrics, such as recall times or recognition accuracy, they can be reliable – although, of course, the proxy metric must still be related to the researchers' goals. Where user studies attempt to evaluate directly the artistic or aesthetic value of non-photorealistic images, I think the effort is misguided at best.

4 User Studies

With respect to research in NPR, there are four chief roles that user studies can play. First, studies can measure the usability and usefulness of tools for image creation and editing. There are many approaches to such studies, both quantitative and qualitative, and the field of HCI offers a huge amount of expertise in designing and executing such studies. Second, studies can measure the effectiveness of images rendered for a specific purpose: for example, making maps comprehensible, or drawing the eye towards designated points of interest. This relates to aesthetics indirectly if at all; in this case, the concerns are strictly practical, i.e., whether the rendering style facilitates the stated task. Third, we can employ user studies in an exploratory way, attempting to elicit responses to stimuli or to learn something about how stimuli are perceived. Questions such as "where do people draw lines?" [Cole et al. 2008] can be investigated by user studies. Such studies can be fairly structured, as in the line-drawing study, or use relatively unstructured approaches such as pile-sorting [Isenberg et al. 2006]. Fourth, user studies can attempt to compare the aesthetics of images directly, for example by having users vote on the relative aesthetics of image pairs, as in the humor site kittenwar.com [2014]; a slightly more serious example comes from the inscribed curve comparisons of Wyvill et al. [2012].

Regrettably, researchers sometimes attempt to use studies to validate results. That is, there is a predetermined conclusion – some variation of "our rendering method makes nice images" – and the goal of the supposed experiment is to demonstrate that the conclusion is true. Conversely, a study undertaken in a scientific mode of inquiry seeks to learn something and information is gained no matter what the outcome. Studies for validation are arguably unscientific in intent, but that would not much matter if they were undertaken in a rigorous way, with sober consideration given to the best way to extract the underlying facts of the case. Often, they are not.

A straw-man questionnaire is shown in Figure 2. It would be nice to believe that this is strictly a parody, but it is uncomfortably close to some real questionnaires. It is not so much designed to investigate the participants' true beliefs as to allow them to express a mild positive judgement, or at least to make responses that could be construed as expressing a positive judgement. As Hertzmann points out, authors sometimes prepare user studies and reviewers request them, even when the information they add to the paper is zero or even negative, shutting down inquiry. Hertzmann is careful to state that his evidence that this happens is strictly anecdotal, but it lines up so well with my own experience that I am prepared to believe that the phenomenon is fairly widespread, albeit (I hope) uncommon. I have exaggerated the questions in the figure for effect, but I have seen almost equally uninformative questions in papers submitted for publication.



Reply to the following questions with respect to the image above. Answer using the following scale: 1=Strongly Disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly Agree.

...
x. I am sophisticated enough to see the beauty in the image.
xi. The image is aesthetically appealing.
...
xiv. The image reminds me of modern art.
xv. The image looks more like a Picasso than a Rembrandt.

Figure 2: *A parody of a questionnaire*

I do not assume bad faith on the part of authors seeking to validate their results by way of user studies. Such studies are a natural reaction to a perceived need for objective evaluation. Still, user studies aimed at validation are potentially damaging to the field. They waste effort, provide a false sense of security by supporting a pretense that aesthetic results have been evaluated scientifically, and inhibit reflection and introspection from authors and readers both. In the following, I expand on the chief reasons why I think user studies should not be used to validate image output.

Participants may be "good subjects". The so-called "good-subject effect" [Nichols and Maner 2008] occurs when excessively prosocial experimental subjects are aware of the hypothesis and attempt to help the experimenter by confirming it. In the case where the hypothesis is "our method is great" and the question is "this image is great, do you agree?" it takes no special insight from the participant to guess the experimenters' intent. The good-subject effect can be mitigated by careful study design: for example, by taking care to disguise the hypothesis or even to deceive participants about the nature of the experiment. However, studies in NPR are usually not very carefully designed, because of the next problem.

We don't take user studies seriously. Few in the NPR community have significant expertise in HCI. Studies are often an afterthought, a box to be checked, and hence are designed and conducted in a superficial way. Recall that I am strictly referring to studies undertaken to confirm the quality of results already believed to be good by the experimenter; if, counterfactually, the experimenter did not believe the results were very good, the study would not have taken place at all.

One might contend that the solution is for NPR practitioners to become experts in study design. However, there is an opportunity cost to gaining expertise in any field, and there are many fields competing for researchers' attention, including areas of mathematics, physics, perception, art history, and computer graphics more broadly. This list is far from exhaustive. I predict with high confidence that not all NPR practitioners will concentrate on learning about study design. Moreover, expertise in study design is a necessary but not sufficient condition for evaluating using a human-subject experiment, as I discuss next.

Studies demand significant time and effort. Even if we had the desire and expertise to create a meaningful study, we still might not want to. Greenberg and Buxton [2008] wrote at some length cautioning the HCI community against needless formal usability evaluations. The path from concept to execution of a user study, involving securing approval from the institutional review board, designing the experiment and finding participants, conducting the study, and analyzing the data afterwards, can easily take weeks or even months.

Once the study begins, the algorithms and results under study should be considered frozen. If improvements are found, either they must be presented without evaluation or the study redone. Even if we are willing to consider redoing the portions of the study relating to an improvement, it cannot be done lightly. Studies have high fixed overhead: we cannot typically redo just 10% of a study at 10% of the cost. Paradoxically, rerunning a study can be more challenging than running it the first time, depending on the depth of the participant pool: we should not reuse participants, and as those easiest to recruit are spent, the difficulty of finding more rises. (Doing large-scale online studies using platforms like Amazon's Mechanical Turk remedies this problem, but exacerbates the problem of participant disengagement discussed next.) Overall, the burden of studies drains effort from other aspects of the research program, disproportionately so given our general lack of expertise in the area. The time spent working on the study could have been used to investigate other questions or to press harder on the questions already under scrutiny.

User studies cannot substitute for expert judgement. Subjects may be hasty and careless; they may be asked extremely demanding questions that they do not think carefully about. They may not follow instructions. In fact, experimenters must beware of giving the subjects too much direction, lest they bias the results. At conference presentations, I have sometimes witnessed an exchange equivalent to the following: *Q. Did you define 'aesthetics' for your participants? A. No, we left it to their interpretation.* I believe this is the proper protocol but we must be quite cautious of over-interpreting surveys conducted in this way.

Kashdan [2014] summarizes the problem thus: "if you ask a question, people will answer it, even if the question doesn't make sense or is far beyond their computational capacities." In a survey relating to subjective reactions to images, responses will vary depending on the subjects' understandings of what is being asked, which is typically not specified with any great precision. Subjects may judge casually and use irrelevant criteria. The researchers themselves, however, can judge carefully and deliberately. They will have an excellent grasp of the issues and can define their terms for the reader. They can also expose details of their thinking: providing such details will allow readers to reproduce their reasoning. In a quantitative user study, the thinking of the subjects is opaque.

There is no way to independently verify the results short of another study. The data might be available, but flaws in the study design – and it is a rare design that is entirely flawless! – call the data into question, and we typically cannot dig into anything behind the raw data. We cannot, for example, probe the participants' intentions, ask them collectively or individually what they might have meant by rating this image a '4' and this other a '3'. The results take on a solidity which may not be warranted given the unstated assumptions and myriad influences confounding the subjects' responses. Qualitative studies using processes such as structured interviews are less susceptible to this criticism, note; as studies become more qualitative, they begin to resemble the principled subjective evaluation I am advocating.

In contrast to quantitative user studies, principled subjective evaluation allows readers to examine the assumptions and reasoning behind the conclusions, since the conclusions are to be justified by argument. Where the reader disagrees with a conclusion, in principle the source of disagreement can be identified; if this occurs prior to publication, the argument or the algorithm can be amended. Disputing the outcome of a user study is unhelpful: at most, a reviewer can point out flaws in the study and the study can be redone. Redoing a study is costly in time and effort and may not touch on the real underlying issue. In the case of a dispute over a conclusion justified by argument, the argument may have strong points as well as weak points and some of the reasoning can be salvaged.

When intended to validate image quality, user studies are epistemologically suspect. The researchers must already believe that their results are good; they would not have attempted the study otherwise. Thus, the study provides no new information to the researchers. Few researchers in NPR would set aside their algorithm on the grounds that the user study was negative or indeterminate. To a general reader, the outcome of the study has very little usefulness: the reader will examine the results and form an independent judgement, setting aside the evidence provided by the user study. The role of the study is transactional, a ceremonial exercise for the authors and reviewers en route to getting the paper published. Once published, though, the inscrutable numerical summary of user responses loses its purpose; it will change the opinion of no reader. It will be much better for the reader to see the researchers' explanation of the good and bad aspects of the work, the omissions deliberate and otherwise, and the original objectives described in detail. Information of this sort should be presented anyway; let's use it for evaluation directly.

5 On the Applicability of Authorial Subjective Evaluation

I have argued in favor of a thorough, systematic subjective approach to evaluating results: or rather, to evaluating a certain kind of result. The intent is to apply this evaluation methodology when the output images are supposed to represent a new style or an improved attempt to render synthetic images in a traditional style. They should have the goal of "being art" rather than any more specific goal, such as information visualization. Not all research efforts in NPR fulfill these criteria.

Authorial subjective evaluation of output images is less useful when any one of the following apply. It is never entirely useless to evaluate subjectively along the lines I described; in particular, the authors' examination of the output images and discussion of the perceived strengths and weaknesses will always be welcome. However, when an objective or quantitative evaluation method is suitable, authorial evaluation will be explanatory and supplemental rather than central.

✗ **When ground truth is available.** Distance from ground truth is an objective metric that is always appropriate, though the proper distance metric may not be obvious. We usually do not have ground truth when working in NPR, but it may be available for certain

problems or applications. Certainly, some related problems such as noise removal allow evaluation by comparison with ground truth.

✗ **When there is a clear task.** If there is a task for a user, there are probably some measurable outcomes, so a quantitative user study is a suitable evaluation mechanism. Even if the outcomes are not quantifiable – the task is "write a moving poem", say – qualitative methods from HCI can be employed. Note that the study is done to evaluate the tool and how well the user was able to navigate the task; the result of the task (whether the poem was moving or not) should still not be judged by means of a user study.

✗ **When the image style is not reducible to distinct, orthogonal elements.** When the style depends on a particular form of cooperation between multiple components, it becomes quite difficult to separate the features for systematic independent evaluation. This is not fatal, and the stages of authorial subjective evaluation can still be attempted, but it may be less informative owing to the interdependence.

Conversely, authorial subjective evaluation is most readily applicable when all or most of the following considerations are met.

✓ **When the overall goal is to reproduce a distinct style for which examples are available.** This is the scenario in which the analysis of the style is easiest and the evaluation of the replication of the different elements will be most effective.

✓ **When there is no relevant objective function.** Usually there is no suitable objective function, but when there is, the objective function can be used for evaluation. Demonstrating the relevance of the function can still be done using subjective evaluation, though.

✓ **When the goal is to make images that interest an audience.** If there is a more specific purpose for the image synthesis process, such as reducing file size or removing noise, evaluation can be done according to whether the purpose was met or not. Further, more well-defined purposes are more easily captured by an objective function.

Before concluding, I will offer some final thoughts on the general applicability of authorial subjective evaluation in NPR.

Authorial subjective evaluation is for evaluating images. It does not seek to evaluate tools or algorithms, but the resulting images only. Tools are usually task-oriented and can be evaluated on usability grounds. Algorithms can be evaluated on the basis of standard norms such as elegance, novelty, robustness, and resource demands (runtime, memory).

Authorial subjective evaluation is not objective. An obvious point, but it means that we have moved towards the norms of art and design, rather than those of science. This is not as fatal as it may seem: art and design have as long a history as science, with many successes to point to. Computer graphics has always charted a course between art and science; it is a pragmatic discipline, interested in what works, often with respect to artistic and commercial considerations.

Subjective image evaluation is incomplete. It does not speak to robustness, effort, practicality, or other properties of a method that we may care about. It is strictly about evaluating images, independent of the underlying process to the extent this is possible. Having discussed image quality, the authors must also discuss other aspects of the method; fortunately, these are generally more susceptible to objective measurement.

6 Conclusion

I advocate a lightweight, subjective evaluation scheme intended for non-photorealistic images, and in particular, images that were created with no particular application, only a desire to intrigue and delight audiences. Objective evaluation of this kind of work is difficult owing to the lack of a clear problem statement. While user studies appear to offer a solution, previous position papers [Hertzmann 2010; Hall and Lehmann 2013] cautioned against them. Likewise, I also argued against applying user studies when evaluation of this sort of result is needed. User studies can be applied sensibly when there is a tool or task, but have little to offer when we have an audience rather than a user.

The principled subjective evaluation I favor hearkens back to the earliest days of NPR. Sadly, even the basic prerequisite to evaluation – a clear statement of the objectives – is sometimes neglected in recent work. The systematic approach taken by Curtis et al. [1997] nicely lays out its objectives and allows the reader to evaluate each component separately. The approach of basing evaluation primarily on visual inspection is still widely used; the main purpose of this position paper is to argue for a more rigourous application of this norm, supplemented by clear statements from the researchers about their goals and specific details about how the goals were met or were not met. Authorial subjective evaluation need not be the sole evaluation mechanism, and it is not always an appropriate approach to evaluation, but when it is relevant it should be applied even if augmented by another evaluation process.

Acknowledgements

A huge thanks to Robert Biddle for several discussion sessions in which he helped me refine my ideas. Thanks also to the anonymous referees, whose insightful suggestions helped improve the paper in many ways. Finally, thanks to NSERC and Carleton University for their financial support of this work.

References

COLE, F., GOLOVINSKIY, A., LIMPAECHER, A., BARROS, H. S., FINKELSTEIN, A., FUNKHOUSER, T., AND RUSINKIEWICZ, S. 2008. Where do people draw lines? *ACM Trans. Graph. 27*, 3 (Aug.), 88:1–88:11.

CURTIS, C. J., ANDERSON, S. E., SEIMS, J. E., FLEISCHER, K. W., AND SALESIN, D. H. 1997. Computer-generated watercolor. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, SIGGRAPH '97, 421–430.

DECARLO, D., AND SANTELLA, A. 2002. Stylization and abstraction of photographs. *ACM Trans. Graph. 21*, 3 (July), 769–776.

GOOCH, A. A., LONG, J., JI, L., ESTEY, A., AND GOOCH, B. S. 2010. Viewing progress in non-photorealistic rendering through Heinlein's lens. In *Proceedings of the 8th International Symposium on Non-Photorealistic Animation and Rendering*, ACM, New York, NY, USA, NPAR '10, 165–171.

GREENBERG, S., AND BUXTON, B. 2008. Usability evaluation considered harmful (some of the time). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, New York, NY, USA, CHI '08, 111–120.

HALL, P., AND LEHMANN, A.-S. 2013. Don't measure – appreciate! NPR seen through the prism of art history. In *Image and*

video-based artistic stylisation, P. Rosin and J. Collomosse, Eds. Springer, Springer-Verlag London, 333–351.

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

ISENBERG, T., NEUMANN, P., CARPENDALE, S., SOUSA, M. C., AND JORGE, J. A. 2006. Non-photorealistic rendering in context: An observational study. In *Proceedings of the 4th International Symposium on Non-photorealistic Animation and Rendering*, ACM, New York, NY, USA, NPAR '06, 115–126.

ISENBERG, T., ISENBERG, P., CHEN, J., SEDLMAIR, M., AND MOLLER, T. 2013. A systematic review on the practice of evaluating visualization. *IEEE Transactions on Visualization and Computer Graphics 19*, 12, 2818–2827.

ISENBERG, T. 2013. Evaluating and validating non-photorealistic and illustrative rendering. In *Image and video-based artistic stylisation*, P. Rosin and J. Collomosse, Eds. Springer, Springer-Verlag London, 311–331.

KASHDAN, T., 2014. 5 psychological studies that require a second look. http://www.psychologytoday.com/blog/curious/201402/5-psychological-studies-require-second-look.

KITTENWAR, 2014. Kittenwar! May the cutest kitten win. http://www.kittenwar.com.

KOSARA, R., DRURY, F., HOLMQUIST, L. E., AND LAIDLAW, D. H. 2008. Visualization criticism. *IEEE Computer Graphics and Applications 28*, 3, 13–15.

KUSHNER, T. K. 1983. *The anatomy of art: problems in the description and evaluation of works of art*. Warren H Green Inc.

KYPRIANIDIS, J. E. 2011. Image and video abstraction by multi-scale anisotropic Kuwahara filtering. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Non-Photorealistic Animation and Rendering*, ACM, New York, NY, USA, NPAR '11, 55–64.

LEE, S., OLSEN, S. C., AND GOOCH, B. 2006. Interactive 3d fluid jet painting. In *Proceedings of the 4th International Symposium on Non-photorealistic Animation and Rendering*, ACM, New York, NY, USA, NPAR '06, 97–104.

MCCLOUD, S. 1993. *Understanding comics*. Kitchen Sink Press.

NICHOLS, A. L., AND MANER, J. K. 2008. The good subject effect: investigating participant demand characteristics. *Journal of General Psychology 135*, 151–165.

PANG, W.-M., QU, Y., WONG, T.-T., COHEN-OR, D., AND HENG, P.-A. 2008. Structure-aware halftoning. In *ACM SIGGRAPH 2008 Papers*, ACM, New York, NY, USA, SIGGRAPH '08, 89:1–89:8.

PEPPER, S. C. 1949. *Principles of art appreciation*. Harcourt, Brace, and Company.

SZELISKI, R., ZABIH, R., SCHARSTEIN, D., VEKSLER, O., KOLMOGOROV, V., AGARWALA, A., TAPPEN, M. F., AND ROTHER, C. 2006. A comparative study of energy minimization methods for Markov random fields. In *ECCV (2)*, 16–29.

TRAN, N. 1999. Generating photomosaics: An empirical study. In *Proceedings of the 1999 ACM Symposium on Applied Computing*, ACM, New York, NY, USA, SAC '99, 105–109.

WINNEMÖLLER, H., OLSEN, S. C., AND GOOCH, B. 2006. Real-time video abstraction. *ACM Trans. Graph. 25*, 3 (July), 1221–1226.

WYVILL, B., KRY, P. G., SEIDEL, R., AND MOULD, D. 2012. Determining an aesthetic inscribed curve. In *Proceedings of the Eighth Annual Symposium on Computational Aesthetics in Graphics, Visualization, and Imaging*, Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, CAe '12, 63–70.