Chapter 15
Evaluating and Validating Non-Photorealistic and Illustrative Rendering

Tobias Isenberg

Abstract In many areas of non-photorealistic and illustrative rendering, considerable progress has been made toward synthesizing traditional artistic and illustrative techniques. However, evaluation and validation of such images have only been attempted relatively recently. This chapter surveys evaluation approaches that have been applied successfully in non-photorealistic and illustrative rendering. It provides an overview over different evaluation approaches including qualitative and quantitative techniques and gives examples for how to approach evaluation in the NPR context. Collectively, the described techniques do not only answer the question of whether an NPR technique is able to replicate a traditional technique successfully but also what implications the use of NPR techniques has and what people think about different NPR techniques as compared to traditional depictions.

15.1 Introduction

With non-photorealistic, artistic, and illustrative rendering (which is collectively being called NPR in this book) having developed into a mature field over the last two to three decades, researchers have begun to question the validity, usefulness, appropriateness, and acceptance of the large variety of different techniques that have been created [13]. This chapter aims to survey the different evaluation and validation techniques that have been employed within NPR to provide inspiration for future work and to encourage the use of evaluation techniques in the field. For the purpose of this chapter on evaluation, however, we treat the domain of NPR a bit more broadly than only the stylization of images and video as in the rest of this book: we incorporate all NPR approaches in the discussion, including those that use 3D scenes as input as well as methods for illustrative visualization.

Tobias Isenberg
INRIA Saclay, Orsay, France
e-mail: tobias@isenberg.cc

To be able to discuss specific NPR evaluation strategies, however, we need to start by thinking about what it is that we want to or need to learn. In fact, there are many different questions one may ask about the field of NPR as a whole or about individual techniques. Hertzmann [26], for example, talks about evaluating human aesthetics and the question of how people respond to NPR, while Salesin [47] mentioned the *NPR Turing test* as one of his seven grand challenges for NPR in 2002 (recently revisited by Gooch et al. [16]): Can we render images using NPR that a normal person is no longer able to distinguish from hand-made ones? While answering these questions is certainly a worthwhile endeavor, the potential for evaluation within NPR is much larger. One may ask, for example, the following questions:

- Why do we want to or need to use NPR in the first place?
- What are appropriate goals for NPR?
- Is a given approach/technique/application accepted by its intended users, does it serve the intended purpose?
- By which mechanism/principle does NPR imagery assist a given goal, and how can we take advantage of such mechanism/principle?
- What do people think about NPR imagery or how do they respond to it?
- What emotions or (potentially) unconscious reactions can/does NPR imagery invoke in viewers?
- How do NPR images compare to hand-made drawings/paintings/illustrations?

Each of these points, in turn, cover a broad range of more specific questions (see also [16, 26, 47]). To be able to discuss NPR evaluation in a more systematic way we, therefore, group these questions roughly into three major areas:

1. the question of providing a general motivation for the use of NPR techniques (Section 15.2),
2. the question of understanding how NPR techniques support a specific purpose (Section 15.3), and
3. the question of comparing hand-made images with computer-generated (non-photorealistic) ones (Section 15.4).

Before we turn to discussing these three main questions, however, we need to briefly touch on study methodologies in general. We do this, in particular, because there is a danger of selecting a wrong study methodology [20, 26] or misinterpreting the results [5, 32]. While a more comprehensive overview of study methodology can be found elsewhere (e.g., [11, 34]), a useful short overview is given by Carpendale [4] for the domain of information visualization but which similarly applies to the study of NPR. Generally, there are two major types of evaluation methodologies that can be employed: *quantitative evaluation* which focuses on hypotheses, measurable variables in controlled experiments, and a statistical analysis of the results and *qualitative evaluation* which tries to gain a richer understanding of the subject matter by taking a more holistic approach and which uses techniques like observation and interviewing [4]. Both general techniques as well as combined approaches can be and have been applied to NPR evaluation, the specific type of methodology depending on the questions that one is asking. For example, the question of what effect a stylized depiction vs. photograph have on learning and recognizing [19] is rightfully studied with a quantitative technique, while the question of what people think about hand-drawn illustrations vs. computer-generated visualizations [30] is better studied with a qualitative approach. We illustrate the various methodologies further in the discussion of the individual techniques below.

15.2 Providing a General Motivation for NPR

While the ability to create images in a specific artistic or illustrative style can be motivation enough for the development and application of NPR techniques, we can also examine people's reactions to seeing NPR visuals to better understand why it makes sense to use NPR in the first place. This insight in the general motivation for the use of NPR can then inform the design of new techniques as well as their practical application.

Such an early example of "assessing the effect of non-photorealistic[ally] rendered images" was presented by Schumann et al. [50] in 1996. They were motivated by the continued use of hand drawing in the domains of architecture and CAD [52] and examined the effect that a sketchy rendering style (as opposed to 'normal' shading and regular CAD plots) has on the communicative goals during the development of architectural designs. To study this effect, the authors started from three hypotheses: (1) that sketched depiction styles are preferred to CAD plots and shaded images for presenting early drafts of architectural designs, (2) that sketches perform better in communicating affective and motivational aspects of an image, while CAD plots and shaded images perform better in cognitive aspects, and (3) that sketches stimulate viewers to participate in an active discussion and development of a design, more than shaded images. To examine these hypotheses, Schumann et al. [50] used a questionnaire-based approach that both asked for quantitative ratings (selection of an image that is preferred for a given task or responses on a 5-point Likert scale) and for qualitative feedback. The questionnaires were sent to 150 architects and architectural students, 54 of which (36%) returned it. Based on these responses the authors analyzed their three hypotheses.

The results showed that of the people who regularly use CAD tools (67% of the responses), 53% would use the NPR sketch to present an early draft, while only 33% would use a CAD plot for this purpose and only 22% would use a shaded image. In contrast, only 8% would use the NPR sketch for a final presentation, while a CAD plot would be used by 50% and the shaded image by 42%. These results confirm the first hypothesis and, thus, show that stylistic depictions can be used to indicate the stage of a design process—a fact that has since then been used, for example, in the domain of sketch-based interaction and modeling (e.g., [27, 49]).

To analyze the second hypothesis the participants were asked to assess the impression that the three different images have on them in more detail. The authors used a classification scheme from the psychology literature and asked the participants to rate each image according to criteria from a cognitive group, an affective group, and a motivational group. They found that the NPR sketch was rated significantly higher/better on affective and motivational criteria, while the CAD plot was rated significantly higher/better on cognitive criteria. This result indicates the potential for stylistic depiction to evoke emotion and to stimulate active involvement, a question that was further examined with respect to the third hypothesis.

To analyze this involvement in a design process, the CAD-using participants were asked how they would communicate design changes using either an NPR sketch or a shaded image: (a) using verbal descriptions, (b) using gestures or pointing, (c) by drawing onto a separate sheet of paper, or (d) by drawing into the presented image. The only statistically significant difference that was found between the NPR sketch and the shaded image was that participants are much more likely to draw into the NPR sketch (69%) than into the shaded image (33%), confirming the authors' third hypothesis. This means that style of a rendering can have an effect on how willing somebody is to interact with the depiction, and the authors suggest that the stylization leaves more room for interpretation with respect to the exact design.

This effect of stylistic imagery on people—in other words whether and how people are affected by NPR—was also examined by Duke et al. [10] and Halper et al. [21, 22] who describe a motivation for employing NPR styles based higher-level psychological principles. For example, Halper et al. [21, 22] discuss the effect of figure-ground segregation as supported by NPR elements such as silhouettes and feature lines. Their study looked at whether people would rather select objects from an image that were depicted with an abstracting style consisting of cartoon shading and silhouette or objects shown in a more detailed, oil-painting style. Their results suggest that the rendering style had an effect on which objects people selected, with participants tending to select two or more objects depicted in the cartoon-style.

A second evaluation by Halper et al. looked at people's social perception and judgment. They presented participants with simple line drawings of scenes with previously established social connotations about safety and danger, such as a house (typically associated to be safe) and a group of trees (typically considered to be less safe). They then depicted the house in a zig-zaggy and the trees in a rounded style, which changed participants' behavior to no longer associate the house with safety. A simple comparison of the same object rendered in different styles supported these findings, also associating certain (zig-zaggy) line styles with danger.

Finally, Halper et al. examined aspects of environmental psychology and people's participation and interaction in environments. In particular, they studied how the level of detail in a rendering affects people's behavior. They provided images with two paths, one depicted in more and another in less detail. They found that the amount of detail has an effect on the choice of path people make—participants preferring the more detailed path over the one with less detail.

Based on these and other experiments, Duke et al. [10] explain people's behavior using the concept of *invariants* from perception, which describes a property common to or shared by a range of entities of behavior. Duke et al. argue that the stylistic differences (i.e., NPR styles) and their stylistic invariants lead to specific behavior in the experiments due to latent, implicational knowledge and that this may lead to higher-level cognitive interpretations common across a range of people (affected by their culture and language). This insight can then be used, as suggested by Duke et al., to associate depicted objects with emotions or to guide people's engagement and attention, for example in computer games and virtual environments.

These two aspects—the effect of NPR on emotion and the use of NPR to guide attention—which both form a strong motivation for the use of NPR in practice have been studied in more detail by two other teams of authors. The first aspect, the effect that NPR imagery has on people's emotions, has recently been studied in detail by Mandryk et al. [41] and Mould et al. [44] who looked at how the stylization of photographs changed people's emotional response to the images. They selected 18 images covering a wide range of topics from the IAPS image database (created specifically as affective stimuli and with known emotional content), and examined the emotional response of 42 participants. They measured the emotional response of their participants with respect to an established dimensional scheme of emotion (using a 5×3 pictorial scale that allowed participants to analyze and report their emotional state) including the dimensions valence, arousal, dominance, and aesthetics. For that purpose they compared the emotional response of the original images (whose previous rating for affective content was known) with those of the result of five image-based NPR styles and those of two blurred versions.

The result of Mandryk et al.'s [41] and Mould et al.'s [44] analysis was that all NPR techniques significantly shifted their participants' reported experiences of valence (pleasure/positive or displeasure/negative of a feeling) and arousal (energy/activation of a feeling) to the neutral rating, thus reducing the strength of the emotion, but never eliminating the emotion completely. They found that some techniques preserve the emotions better than others, but that the effect might be attributed to the amount of detail that was preserved by a given technique. It is interesting that this muting of emotion to some degree stands in contrast to the observations of Duke et al. [10] and Halper et al. [21, 22], but this effect can probably be explained based on the different types of stylization employed by both evaluations: Mandryk et al. [41] and Mould et al. [44] examined image-based (i.e., mostly space-filling) techniques in which the style's amplitude (as measured in tone or color) is reduced by the NPR technique from the original photograph due to the introduced abstraction, while Duke et al. and Halper et al.'s analysis of emotion (social perception and judgment) was based on line drawings in which the strong emotion (fear) was caused by a high-amplitude zig-zaggy style.

Another recent study adds to these mixed results: Seifi et al. [51] looked at what effect color palettes have on the perception of emotion in painterly rendered faces. They used color palettes designed to enhance certain emotions (joy, surprise, anger, and fear) and examined their effect for still images and animations. Seifi et al. found that sometimes the perceived emotion is emphasized if the palette matches the face's expression, while non-matching palettes dampen the perceived emotion. However, they also report about a general damping effect for some emotions, and that even sometimes the perceived emotion is damped further when the matching palette is used than for non-matching palettes (e.g., fear in the animated scenario).

The second general aspect to be affected in a controlled manner by NPR styles as suggested by Duke et al. [10] and also previously mentioned by Strothotte et al. [52]—the guiding of people's attention—was studied in detail by Santella and DeCarlo [48]. Their goal was, based on eye tracking data, to understand the effectiveness *meaningful abstraction* (i.e., directed removal of detail) with the intent of guiding a person's attention. For that purpose Santella and DeCarlo created four types of abstraction of an input photograph: one using constant abstraction with a high amount of detail, one using constant abstraction with a low amount of detail, one with adaptive abstraction in which the detail points are based on image saliency, and one with adaptive abstraction in which the detail points are based on a person's fixation points for the original photograph. The authors then used their eye tracker to analyze where 74 study participants fixated when looking at one of the five different versions (original and four abstracted versions) of 50 input images (using a between-participants design).

The study results showed that the local treatment of abstraction does have an effect on where people look in an image, the salience-based and fixation-based adaptive abstractions receiving fewer fixation clusters than the other images. The analysis of the distance of the fixation clusters to the detail points also showed that these are smaller for both adaptive abstractions, leading to a concentration of the visual interest. Moreover, the authors also argue that the distance from a detail point to a cluster is consistently smaller for the eye tracking condition than for the salience condition, suggesting that eye tracking points were more closely examined, i.e., that there seems to be less interest in salience points. These results provide evidence for the previously discussed hypothesis that the control of the amount of detail as it is possible with NPR styles can be used to guide people's attention.

Related to the use of NPR for guiding attention is also the issue of whether or not the abstraction introduced by NPR techniques has a positive effect on the ability to recognize and to memorize objects. This question was examined by Gooch et al. [18, 19] specifically for face illustrations, but similar to the previous studies their work provides a motivation for NPR in general. In their psychophysical study the authors compare photographs of faces with line illustrations as well as with line-based caricatures of the same faces. Specifically, Gooch et al. used controlled, quantitative experiments to measure the speed and accuracy of participants recognizing known faces and learning new faces, using photographs, computer-generated line illustrations, and caricatures produced with computer support.

In a first experiment, Gooch et al. used images created from face pictures of the 12 most familiar out of 20 possible people (colleagues) and asked their 42 participants to recognize them when presented in a random order. Each participant saw only two types of images, either photographs and illustrations, photographs and caricatures, or illustrations and caricatures. The results of this experiment were that participants were slightly faster in naming photographs than caricatures, with the other combinations not showing a significant effect and with the accuracy for all conditions being high (98%)—thus without a speed-for-accuracy trade-off.

To examine the learning of the different types of depictions in a second experiment, Gooch et al. created the same types of images for faces that were unknown to 30 different participants. As before, each participant was shown 12 face pictures, but this time a name was associated with each picture and each participant only saw one type of image (i.e., three groups of 10 participants). Then, the image stack was shuffled and the participants were asked to recall the name previously associated to a face as the stack of images was presented one image at a time. If a name was incorrect, then the participant was corrected. This process (including shuffling) was repeated until all names were recalled correctly. The analysis showed that illustrations were learned more than twice as fast as photographs. While caricatures were about 1.5× as fast as photographs, this difference was not statistically significant.

Interestingly, a similar experiment about the ability to recognize and memorize objects using abstractions vs. real photographs was later conducted by Winnemöller et al. [55], using their real-time video abstraction technique as a foundation (Chapter 5). This meant that they could investigate color images as opposed to the black-and-white images used by Gooch et al. [18, 19]. Like Gooch et al., Winnemöller et al. used faces for the recognition task (using celebrity images) but employed arbitrary scenes for the memory task in a memory card game setting. Overall, their results supported the findings by Gooch et al., but in contrast also found that recognition was significantly faster with the abstracted images as opposed to the real photographs. However, this result does not generalize to all NPR depiction styles—Zhao and Zhu [56], for example, showed that objects depicted in both painterly rendering and actual paintings are recognized slower than actual photographs.

In summary, the mentioned studies provide evidence for a number of benefits or effects of stylistic depictions created with NPR techniques, and thus motivate the development of as well as the application of NPR approaches in practice. This includes that they can encourage participation in design discussions [50], that they can assist figure-ground segregation, can carry social connotations, and can steer people's interest [10, 21, 22], that their style has an effect on people's emotions [41, 44], that they can be used to guide people's attention [48], and that they can affect how people recognize and memorize depictions of objects [18, 19, 55].

15.3 Understanding How NPR Supports a Specific Purpose

While the studies discussed in Section 15.2 necessarily each use a specific study setting, their findings do provide a motivation for employing NPR styles more generally as we just outlined. However, there have also been a number of evaluations of NPR techniques that, we find, point to more specific usage possibilities because they illustrate how NPR can support a specific purpose or application domain. We first discuss a number of evaluations that examine aspects that relate to human perception of NPR with respect to textures, then examine techniques that support the creation of visualizations using NPR approaches, and finally look at application contexts in the domains of virtual and augmented reality (VR and AR).

15.3.1 Perception of NPR Textures

A large body of NPR work addresses the creation of textures for a variety of application domains. For example, the creation of stippling (dot placement, see Chapter 3) and hatching (line placement) are among the most fundamental NPR techniques, applicable both to the representation of 2D images and the depiction of 3D shapes. Thus it is important to understand how people see and interpret such textures.

To obtain this kind of understanding, Kim et al. [35, 36] looked specifically at the use of textures for the representation of 3D shapes and their impact on people's shape categorization judgments. In this context it is important that a texture supports the perception of 3D shapes, despite them only being depicted as a projection on a 2D image. Specifically, Kim et al. investigated which effect the texture type has on shape perception, using a set of five different types: one-directional hatching, perpendicular cross-hatching, swirly lines, three-directional hatching, and noise; with three additional variations. Based on these types, the authors conducted a controlled study in which they asked their participants to classify surface patches as ellipsoid, cylindrical, saddle-like, or flat as well as to categorize them as convex, concave, both (for saddles), or none (for flat patches). They found that, overall, the texture type does have a significant effect on shape perception, with the perpendicular cross-hatching along the principal directions performing best, confirming a hypothesis formulated earlier [15]. They also found that certain textures and an oblique viewing direction can alleviate problems that arise from orthographic projection.

To examine the situation further, Kim et al. [36] performed a second experiment to concentrate on single-, dual-, and triple-hatching textures. Some of the hatching directions of the new texture set followed the principal directions, while others were turned away from a principal direction by 45 degrees. Using the same experimental procedure as in the first study, Kim et al. found that, surprisingly, the two-directional hatching texture was now outperformed for some shapes by the single-hatching as well as, in particular, the three-directional hatching. The authors speculate that this effect may be due to people inferring non-existent lines from the otherwise regular hatching patterns as well as distances along these non-existent lines to be able to understand and classify the depicted shape as well as or better than with 'normal' two-directional hatching along the principal directions.

While Kim et al. [36] examined the issue of 3D shape classification based on the applied textures, another problem in NPR is the use of textures on 3D shapes for stylized animation. Traditionally, the straightforward process of applying 2D textures to 3D objects led to a number of issues during an animation including popping, sliding, and deformations—leading to the use of 'fractalized' (i.e., self-similar on different levels of scale) NPR textures in such 3D scene animations. Because these fractalized textures are no longer identical to the traditional textures, Bénard et al. [3] conducted an experiment to analyze this perceived change and, based on these results, to derive a quantitative metric for the introduced texture distortion. In the study the authors asked participants to rank pairs of original and fractalized textures (representative of a variety of media replicated by NPR; including, e.g., stippling, hatching, cross-hatching, and paint texture) with respect to how much distortion participants perceived to have been introduced by the fractalization.

The authors statistically analyze the results and find that, between two sets of textures of the same categories, participants seem to have treated them roughly the same way overall. However, it also is apparent that the class of a texture (e.g., "cross-hatching") does not always get classified the same way if different instances of the texture class are used. More interestingly, however, Bénard et al. try to extract a correlation between known image metrics and their empirical results to find a model that is able to predict a potential perceived dissimilarity of the fractalization for a new texture instance. They find that deriving the average co-occurrence error between the local gray-level co-occurrence statistics of the original and the 'fractalized' version of a texture strongly correlates with the distortion perceived with their participants, and thus suggest it could be used to predict such perceived distortion.

Related to the issue of 'fractalized' NPR textures is the general problem of 2D geometric texture synthesis and the degree of perceived visual similarity between two such synthesized textures. AlMeraj et al. [2] conducted two psychophysical experiments to analyze this question, motivated by the fact that geometric texture synthesis as a sub-domain of NPR is itself based on human perception. In their first experiment, they asked participants to interactively generate a larger dot texture based on a small sample, and then asked the participants both quantitative and qualitative questions to understand their strategies. To analyze the answers to the qualitative questions, the authors used an open coding approach, resulting in the ability to compare responses between participants. From this comprehensive analysis AlMeraj et al. extracted a number of causal attributes that motivated participants' generation styles (dominant visual properties perceived, local themes identifies, and recognition of large spatial structures), a number of strategies for generating geometric arrangements (tiling, structured approach, and random approach), and a number of criteria for evaluating similarity (symmetry, apparent shape, repetition, conformity to apparent rules, and accuracy of copied samples).

To understand the quality of the textures participants had synthesized, AlMeraj et al. conducted a second mixed-method study with a new group of participants to avoid bias. In this study participants were again provided with a sample texture, but this time were shown five synthesized textures. These synthesized textures were derived from the previous experiments (180 textures) which were complemented with 36 computer-generated textures, either using a random or a perfect tiling approach. Participants then had to rank the five shown textures according to their similarity to the shown sample, similar to the approach by Bénard et al. [3]. In addition to extracting the list of criteria used by participants again using a qualitative approach, AlMeraj et al. [2] analyzed the similarities quantitatively, in particular taking the generation strategies (tiling, structured, random) of the first experiment into account. The results show that the textures generated by people using a tiling approach were ranked as "most similar" to the sample textures, likely because participants were able to detect the repeated instances in the larger images. The authors identify an apparent hierarchy in the criteria used for rating similarity: first looking for complete samples, then the identification of themes (small dot arrangements) that are consistently distributed in the synthesized textures, and finally an overall comparison of texture to sample using global mathematical attributes. The overall results are interesting since typically researchers strive for a more structured approach to computer-driven texture synthesis, and also the structured techniques used by participants were also rated higher in AlMeraj et al.'s first study.

15.3.2 Evaluation of Illustrative Visualizations

One specific application area of NPR research is the recently emerged domain of illustrative visualization [45]. In this sub-domain of the general field of visualization it is essential to understand how people see and perceive visuals, thus the evaluation of illustrative visualization plays a particularly important role.

For example, while the previous section examined the evaluation of the perception of NPR textures in general, researchers also have specifically looked at the evaluation of NPR textures in a visualization context. The first approach discussed in this section is closely related to those discussed in Section 15.3.1. This evaluation is particularly interesting because it employs an evaluation strategy otherwise typically used in an artistic context: critique sessions. Jackson et al. [31] and Acevedo et al. [1] report on feedback from expert designers/illustration educators on a number of texture-based visualization techniques to represent 2D vector fields and their properties. The critique by experts as an evaluation strategy promises to provide rich qualitative feedback which can not only suggest which type of technique is better suited for a given purpose but also, in particular, why this is the case. Acevedo et al. [1] also compared the critique-based results by Jackson et al. [31] to those of a previous controlled experiment [37]. They found in their pilot study that the results exhibited the same patterns for both studies, but that the designer critique generally took less time. Both Laidlaw [38] and Keefe et al. [33] describe this critique-based evaluation strategy in the larger context of art-science collaboration, outlining how the evaluation fits into the general visualization design workflow.

In general, NPR and illustrative visualization are both driven by inspiration from artistic practice, leading to a number of artistic visualization techniques. For example, Healey et al. [23, 24] describe a texture-based visualization technique for 2D vector and scalar data that employ techniques from painterly stroke-based rendering [25]. Because the simple inspiration is not sufficient for validating the usefulness of such a technique, they validated their approach using a series of psychophysical experiments. First, they examined whether, in general, people are able to rapidly and accurately identify a group of target brush strokes within a larger stroke set by color or orientation. They found that participants could identify stroke groups better by color than by orientation and that random colors interfere with the identification of orientations. They also concluded that their results indicate that the use of painterly strokes for visualization seems feasible, and thus continued to create one.

Based on these results and also inspired by positive expert feedback, Healey et al. [24] thus conducted a second experiment, this time with their newly created actual 2D visualization of weather data. They examined whether the illustrative visualization could support practical analysis tasks on real-world data and compared their technique to existing (traditional) visualizations. Participants in this controlled experiment were asked to identify which visualization would make it easiest from them to distinguish data aspects such as temperature, precipitation, wind speed, and wind direction; to identify regions in the visualizations with certain combinations of high/low values of the scalar properties, as well as to identify regions with rapid change of temperature. The results of this experiment showed that Healey et al.'s illustrative visualization was as good as or better than the traditional weather visualization for all the tested cases, thus specifically for identifying multi-dimensional patterns in the data (which the study was designed to evaluate).

The techniques examined in Sections 15.3.1 and 15.3.2 thus far aim primarily at two-dimensional NPR and visualization techniques. The domain of scientific visualization, however, primarily examines spatial datasets that are defined in three dimensions, and illustrative visualization approaches have also been developed for this purpose. Consequently, researchers are interested in validating such techniques, for example in the context of medical visualization. Tietjen et al. [53], for instance, looked at the domain of surgery planning and education and, specifically, at the different visualization of detail and context of the human anatomy. Their hybrid visualization technique combined traditional shading techniques (usually for focus objects) with volumetric rendering and NPR line rendering (for both near-focus and context objects), and the authors were interested in how the use of different visualization techniques for depicting, in particular, context objects would support tasks both for medical practitioners and for laypeople.

To examine these questions, the authors employed a questionnaire-based evaluation methodology and distributed the questionnaires to surgeons and participants without a medical education. Each page of the questionnaire showed two different visualizations, one of which needed to be chosen based on personal preference, questions with respect to the usefulness of the visualization for specific tasks were asked (using Likert-scale ratings), and the visualization particularly suited for surgery education needed to be identified. Using this approach, Tietjen et al. compared one specific visualization (their reference) to all other variants they tested, and also conducted comparisons of some other combinations for cross-validation. Based on this approach, the authors conclude that their technique was considered to be appropriate by most of the participating surgeons, and that these surgeons tend to prefer little context information as long as context is present. Laypeople favored images in which the context was shown with colored silhouettes or with silhouettes combined with additional surface shading.

15.3.3 *Perception of NPR in the Context of VR/AR and Immersion*

As we have seen in Sections 15.3.1 and 15.3.2, human perception plays an important role in the evaluation of NPR results. It is particularly important to understand how perception affects viewers if we use NPR in contexts that augment or completely replace the ‘normal’ reality as we experience it every day—i.e., in fields like virtual and augmented reality (VR and AR). While to date there have not been that many approaches to apply NPR in such a context, there have already been some noteworthy evaluations of VR/AR NPR settings.

An early example was presented by Gooch and Willemsen [17] who tried to answer the question how the perception of space and distances in a VR context if the normal (photorealistic) environment was replaced by an NPR one. Specifically, they created a model of their physical lab environment to be able to create a black-and-white NPR version of it using silhouettes and feature lines (creases). With the help of a tracked head-mounted display (HMD) setup they were then able to present participants with the NPR environment as well as with the real world (no HMD). In Gooch and Willemsen’s controlled experiment (within-participants design), the participants were shown a target shape a certain distance away from their position (in either the NPR or real-world condition), were able to look around (turn the head but not move it), and then were asked to walk blind-folded up to the point where they had perceived the object—the distance of which was recorded.

The quantitative analysis of the recorded walked distances showed that, in the NPR condition, participants walked 66% of the distance to the object, while in the real world they would walk 97% of the distance on average. While the difference of walked distances in the NPR condition from the real world seems to be an indication for the inappropriateness of using NPR within VR, the authors argue that the observed over-estimation of NPR-VR corresponds well to how people perceive and interact in ‘normal/traditional’ (i.e., photorealistic) VR environments, thus conclude that NPR-based VR environments are a viable alternative to photorealistic ones.

In fact, when using immersive environments it is not always necessary to decide for either a physical or a virtual world, but it is also possible to combine both in an augmented reality. Such setups add virtual objects to otherwise realistic scenes which are captured through a camera system or see-through glasses. The problem with such setups is that, due to an incomplete knowledge of all environmental influences and the ability to render in a completely photorealistic way, the real and the virtual objects look quite different. To address this discrepancy, people have proposed to use stylized augmented reality which applies stylization to both the real and the virtual parts of the image and thus masks the differences between them. To understand the effectiveness of this approach, Fischer et al. [12] conducted a psychophysical study in which participants were asked to determine if an object shown in a stylized augmented reality setting would be real or virtual.

Based on still images and short video clips (between-participants factor), 18 participants were asked to answer this question for 30 objects (half of which were virtual, the other half real). The results showed that participants were able to correctly determine the type of object in 69% of the cases in the stylized AR style as opposed to 94% in a traditional AR style on average, with the results being consistent between the two groups with still images and video clips. Interestingly, it was more difficult to correctly identify physical objects than virtual objects, but this result was not statistically significant. The authors speculate that these results are due to the lack of compelling 3D models which led to a number of rendering errors which made it easier for participants to tell that virtual objects were, in fact, virtual. However, the authors also conclude that their experiment showed stylized AR to be successful in solving the discrepancy problem as it was more difficult to tell objects apart in the stylized condition.

In dynamic 3D environments—such as the mentioned VR/AR settings—it is often also important to correctly and reliably perceive the shape of objects, which may be assisted by NPR means. In fact, one of the defining and seminal NPR publications [46] argues that adding NPR elements (silhouettes and feature lines) to conventionally rendered objects makes them more comprehensible. To examine if such a correlation really exists, Winnemöller et al. [54] present an experimental framework and psychophysical study that examines the usefulness of a number of shape cues in dynamic environments. Specifically, they analyze how shading, contours, texture, and a mix of shading with contours (which Saito and Takahashi [46] suggest makes shapes comprehensible) affect the recognition of rigidly moving objects.

In the actual experiment, the 21 participants were asked to identify those shapes that moved with other shapes across a touch-sensitive display, using a background similar to the moving foreground objects, that shared a certain shape characteristic. The object depictions and backgrounds were chosen such that they only used contours, only used shading, only used one of two textures, or used a combination of shading and textures. The analysis of the task accuracy showed that the use of only shading lead to the best correct recognition rates, before outlines and textures. Interestingly, the combination of shading with outlines did not perform better than just shading, but worse than it (while still being more accurate than just outlines). While it seems to contradict intuition [46] and studies mentioned earlier [10, 21, 22], Winnemöller et al. attribute this effect to participants feeling that the mixed condition provided too much and, thus, confusing information since also the background shapes were rendered with additional outlines—unlike in the earlier techniques [46] and evaluations [10, 21, 22]. Nevertheless, this example illustrates well that one should not simply base NPR and illustrative rendering design decisions on assumptions but should always validate these assumptions before employing a new technique in a practical application.

15.4 Comparing Hand-Made Images with Computer-Generated Non-Photorealistic Rendering

The final major type of evaluation of NPR techniques to be addressed in this discussion is the question of how NPR imagery compares to human-made images or drawings. This question may initially seem quite straightforward; however, it is not as easy to answer as one may think because it is not clear from the question's general phrasing what we mean by "to compare to." While the NPR Turing test [16, 47]—whether a person is able to distinguish a hand-made from a computer-generated image—that was mentioned in the introduction may be one form of the question, there are also several other possible ways to compare the different types of images.

For example, we may be interested in the question of how drawing patterns differ between hand-made examples and computer-generated styles. This is a question about differences that affect the perceived aesthetics of an image as pointed out by Maciejewski et al. [39] specifically for the area of stipple rendering (see Chapter 3). Maciejewski et al. discuss this difference of aesthetics, overall, in the context of intentional dot placement with several high-level considerations (also discussed in detail by Martín et al. [42]) on the side of hand-drawn stippling—as opposed to a mechanistic stipple placement with many more stipple points, according to simple illumination models and simple stipple shapes, and without the mentioned high-level considerations on the side of the computer-generated stippling.

Maciejewski et al. [40] then proceed to evaluate such differences with respect to the distribution of the stipple dots using statistical methods, specifically by examining the gray-scale textures that characterize the two different styles. For this purpose they employ the gray-level co-occurrence matrix (GLCM) as a tool which measures the frequency in which certain gray levels occur in a given spatial relationship. Based on this matrix they then derive three properties: contrast, energy, and correlation which they use to compare textures from hand-drawn and computer-generated stippling. This analysis unveils a number of differences that Maciejewski et al. initially only hypothesized about: for example, the regular artifacts of stippling based on centroidal Voronoi diagrams (which cause undesired correlation across the textures) as well as similarity of hand-drawn stippling to natural textures. The results also show that, while certain computer stippling techniques that incorporate randomness also exhibit strong correlations with hand-drawn stippling, they still can be easily distinguished from hand drawings due to other regularities. Interestingly, Martín et al. [43] later employed the same statistical evaluation technique to analyze a resolution-dependent halftoning-based stippling with randomness applied to stipple locations but with example-based stipple dots (see Chapter 3, also for example images). The analysis showed that Martín et al. were able to create computer-generated stipple textures whose statistical properties were virtually identical to those of hand-drawn stippling. This suggests that using an example-based approach for NPR as opposed to a purely algorithmic technique may be better able to create results that are less distinguishable from their hand-drawn counterparts.

This observation, however, may not apply to all NPR techniques since some of the primitives used in traditional artistic depiction heavily rely on mathematical principles. One of the best examples for such techniques is the creation of sparse line drawings—lines that consist of silhouettes/contours [29] and feature lines. One question that comes up in this context is "where do people draw lines," [6–8] and how these lines related to the zoo of lines typically employed in NPR; another question is "how well do [such] line drawings depict shape" [6, 9].

To answer the first question, Cole et al. [6–8] conducted a study to compare the line drawings created by artists to depict 3D shapes with NPR-based computer-generated line renderings. For this purpose they conceived an ingenious study setup to satisfy two apparently conflicting constraints: they wanted to (1) allow their participants full freedom in creating line drawings, while (2) at the same time they needed to be able to precisely compare lines created by the artists with lines rendered by NPR algorithms. Cole et al. resolved this conflict by first allowing their participants to draw the 3D shapes freely based on a shaded depiction of the 3D shape, and then in a second step asked them to copy only the drawn lines onto a faint copy of the shaded depiction. This approach resulted in hand-made line drawings that could be compared to computer-generated ones on a pixel basis with a high level of accuracy.

Using this setup, Cole et al. collected input from 29 participating artists or art students, each of whom drew up to twelve 3D shapes, resulting in 208 line drawings in total. To analyze the data, the authors compared the hand-dawn lines both with each other as well as with those lines created by a number of established NPR line techniques including silhouettes/occluding contours, suggestive contours, apparent ridges, image intensity edges, and geometric ridges and valleys.

The analysis showed that artists drew their lines very close to other artists' lines, 75% of the lines being within 1 mm of lines from all other artists. Silhouettes/occluding contours account for most of these similarities, comprising 57% of all lines that were drawn. Other categories of lines from computer graphics/NPR that explain lines the participants draw are large gradients in image intensity as well as object-space feature lines. In fact, all object-space NPR line definitions together account for 81% of the lines drawn by the participants, while each of the category explains some lines that the others do not explain. Overall, the output of all considered line definitions only accounts for 86% of the hand-drawn lines. Cole et al. speculate that the rest could be explained by looking at other local properties, combinations of the existing line definitions, as well as by some higher-level decisions that have to do with what the artists want to communicate or what they think is implied.

Of course, the difference between hand-drawn and computer-generated sparse line drawings is not only interesting from an aesthetic standpoint but also affects how the respective images can be employed for specific purposes. It is particularly important to understand whether it makes a difference to people to use hand-drawn as opposed to computer-generated illustrations of shapes if the people need to correctly perceive, understand, and interpret the 3D shape of the depicted objects. Therefore, Cole et al. [6, 9] conducted a follow-up study based on the data they acquired in their first study to examine the question of shape depiction and perception. For this purpose they employed the established gauge figure protocol to ask people to estimate the perceived orientation of a surface at many points, based on both hand-drawn and computer-generated sparse line drawings (as well as shaded images for comparison). Due to the first experiment's [6–8] setup they also had access to the ground truth in form of the 3D shapes that people illustrated or that were used in generating the NPR images.

Their results show that, for about half of the 3D models they used in the study, the line drawings that performed best were almost as good as the shaded images the authors used for comparison. For other models (e.g., those with organic, smooth, or blobby shapes) viewers had not only more problems understanding the shaded image but also were unable to correctly interpret the line drawings. The study also showed that computer-generated line drawings based on differential properties have the potential to be as effective as hand drawings: for all but one shape there was a computer-generated line drawing that caused lower errors in shape perception than the drawings created by an artists. However, the specific type of lines to be used for a good depiction depends on the specific 3D shape as it was not always the same line type that was responsible for the better result.

The question of which specific primitive to use to depict a given shape not only applies to line renderings of 3D shapes but even more so to pixel art, a unique form of expression arising from early computer depiction [14, 28]. Recently, the question of how to generate such pixel representations from images or vector graphics was examined and evaluated [14, 28]. For example, Inglis and Kaplan [28] create pixel art from vector line drawings and evaluated these by comparing hand-created images with their automatic technique. To achieve this comparison, they first asked their participants to create pixel images for given vector input using a Web tool, and in a second stage asked them to compare these images with each other as well as with synthetically generated ones. Specifically, Inglis and Kaplan asked their participants to compare the images with respect to their visual appeal and with respect to their fidelity. The results showed that the Pixelator technique conceived by Inglis and Kaplan outperformed other computer-generated techniques. The results also showed that, interestingly, people liked Pixelator images better than all groups of human-created images. However, the authors note that this does not mean that Pixelator images outperformed all human-created examples, but instead that their group classification is not a good indicator for how people judge the results. Another interesting result was that images created by people with lots of experience and a high artistic level were not rated very well, likely due to a lot of artistic ‘interpretation’ rather than a faithful depiction as examined by the authors. Here, an evaluation approach that asks for aesthetic judgment [14] may yield different results.

While studies like the ones discussed in this section so far are able to shed light on quantitatively measurable properties of NPR imagery or the suitability of an NPR algorithm, they cannot provide answers to questions about what happens when people look at such images. For example, how do people understand and assess NPR illustrations in general, what do they think about both hand-drawn and computer-generated images, and does a potential difference mean that they would prefer one over the other? Such questions are not easily answered with the more common quantitative evaluation techniques but require a more qualitative approach.

To examine such questions in the context of hand-drawn and computer-generated pen-and-ink illustrations, Isenberg et al. [30] conduced a qualitative, observational study. Specifically, they used an ethnographic approach to avoid biasing people by asking questions in a certain way since any question inherently biases the person asked. The study methodology they chose is an unconstrained pile sorting task which asks participants to ‘sort’ the objects they are given into piles, the specific number and size of the piles being determined by the participant. The objects to be sorted in this case were computer-generated and hand-drawn illustrations of three different objects, each printed on a Letter-sized page. They also had three different types of participants: people with illustration/drawing experiences, NPR researchers, and illustration end users (general university students). Each participant was presented with the pile of illustrations with the images in a random order, and then was asked to form the piles. After that part was completed, the participant was asked to explain what characterized each of the piles in order to understand the reasons for grouping images, and only after this were asked a number of questions about preference, potential use context, and also whether some images/piles looked particularly hand-drawn or computer-generated.

The analysis of the results showed that the different types of participants grouped the images in similar ways and that people generally did group by illustration style and amount of detail. More interestingly, none of the participants constructed piles of the images by whether they thought that an image looked particularly hand-drawn or computer-generated. However, participants were generally able to tell one type from the other with only a few (but consistent) exceptions. Nevertheless, this clear difference did not mean that the participants would like one type better than the other; instead participants liked them for different reasons. For example, participants liked the clarity, precision, three-dimensionality, detail, and—ironically—the realism of the NPR images, while they similarly appreciated the artistic appearance and character of the hand-drawn illustrations. Based on these and other insights from the rich qualitative feedback provided by the participants due to the chosen study methodology, Isenberg et al. provide a number of recommendations and guidelines for future NPR research including to know one's goal, to know one's audience, to explore material depiction and non-realistic models, to avoid patterns and regularities, and to pay close attention to marks and tools.

15.5 Conclusion

The various evaluations of NPR styles and techniques that were introduced in this chapter demonstrate that there are numerous questions that one may want to answer about the produced images. One of the most fundamental ones, however, is the question of the goal of a technique and whether this goal is achieved [30]. One of the obvious goals one may potentially want to strive for is to become indistinguishable from hand-made drawings, paintings, or illustrations. We have seen, however, that the NPR Turing test as proposed by Salesin in 2002 [16, 47] has, to date, not successfully been passed as demonstrated, for instance, by observations by Isenberg et al. [30] or by texture statistics by Maciejewski et al. [40]. Even cases where we as NPR researchers come close (the very few of the NPR images examined by Isenberg et al. [30] that were often thought to be hand-drawn, the stippling distributions examined statistically by Martín et al. [43], or the abstract painterly style of Zhao and Zhu [56]) we can still observe obvious differences, for example on the lack of perceived 'skillfulness' of drawings or the lack of support of higher-level painting/drawing/stippling strategies.

Therefore, the goal of being indistinguishable from artwork is not necessarily the most interesting one for NPR as a field, and thus is also not the most relevant driving force for employing evaluation and validation as part of the research. Instead, the goal of providing general motivations for stylistic rendering as identified in Section 15.2, the need for support of specific goals as discussed in Section 15.3, or the general question of aesthetic judgments (e.g., in Gerstner et al.'s [14] work) may serve as alternative reasons for evaluating and validating NPR algorithms. Nevertheless, comparing one's results to their hand-made counterparts as reviewed in Section 15.4 can also be instructive, but should not only be reduced to an NPR Turing test. In fact, in Chapter 16 of this book Hall and Lehmann look at NPR in the context of traditional artistic depiction and ask the question of how to assess the generated visuals from the perspective of art history. In taking this view, they nicely make the point that an NPR Turing test does not provide any insight on the aesthetic value of the NPR visuals but that the produced images instead have to be appreciated by people—just like traditional artworks.

References

1. Acevedo, D., Laidlaw, D., Drury, F.: Using Visual Design Expertise to Characterize the Effectiveness of 2D Scientific Visualization Methods. In: Proceedings Compendium of IEEE InfoVis and Visualization 2005, pp. 111–112 (2005). DOI 10.1109/VIS.2005.109
2. AlMeraj, Z., Kaplan, C.S., Asente, P., Lank, E.: Towards Ground Truth in Geometric Textures. In: Proc. NPAR, pp. 17–26. ACM, New York (2011). DOI 10.1145/2024676.2024679
3. Bénard, P., Thollot, J., Sillion, F.: Quality Assessment of Fractalized NPR Textures: A Perceptual Objective Metric. In: Proc. APGV, pp. 117–120. ACM, New York (2009). DOI 10.1145/1620993.1621016
4. Carpendale, S.: Evaluating Information Visualizations. In: Information Visualization: Human-Centered Issues and Perspectives, *LNCS*, vol. 4950, pp. 19–45. Springer-Verlag, Berlin/Heidelberg (2008). DOI 10.1007/978-3-540-70956-5_2
5. Cohen, J.: The Earth Is Round ($p < 0.05$). American Psychologist **49**(12), 997–1003 (1994). DOI 10.1037/0003-066X.49.12.997
6. Cole, F.: Line Drawings of 3D Models. Ph.D. thesis, Princeton University (2009)
7. Cole, F., Golovinskiy, A., Limpaecher, A., Barros, H.S., Finkelstein, A., Funkhouser, T., Rusinkiewic, S.: Where Do People Draw Lines? ACM Transactions on Graphics **27**(3), Article No. 88 (2008). DOI 10.1145/1360612.1360687
8. Cole, F., Golovinskiy, A., Limpaecher, A., Barros, H.S., Finkelstein, A., Funkhouser, T., Rusinkiewicz, S.: Where Do People Draw Lines? Communications of the ACM **55**(1), 107–115 (2012). DOI 10.1145/2063176.2063202
9. Cole, F., Sanik, K., DeCarlo, D., Finkelstein, A., Funkhouser, T., Rusinkiewicz, S., Singh, M.: How Well Do Line Drawings Depict Shape? ACM Transactions on Graphics **28**(3), 28(1)–28(9) (2009). DOI 10.1145/1531326.1531334
10. Duke, D.J., Barnard, P.J., Halper, N., Mellin, M.: Rendering and Affect. Computer Graphics Forum **22**(3), 359–368 (2003). DOI 10.1111/1467-8659.00683
11. Field, A., Hole, G.: How to Design and Report Experiments. Sage Publications Ltd., London (2003)
12. Fischer, J., Cunningham, D., Bartz, D., Wallraven, C., Bülthoff, H., Straßer, W.: Measuring the Discernability of Virtual Objects in Conventional and Stylized Augmented Reality. In: Proc. EGVE, pp. 53–61. Eurographics Association, Goslar, Germany (2006). DOI 10.2312/EGVE/EGVE06/053-061
13. Gatzidis, C., Papakonstantinou, S., Brujic-Okretic, V., Baker, S.: Recent Advances in the User Evaluation Methods and Studies of Non-Photorealistic Visualisation and Rendering Tech-

niques. In: Proc. IV, pp. 475–480. IEEE Computer Society, Los Alamitos (2008). DOI 10.1109/IV.2008.75

14. Gerstner, T., DeCarlo, D., Alexa, M., Finkelstein, A., Gingold, Y., Nealen, A.: Pixelated Image Abstraction. In: Proc. NPAR, pp. 29–36. Eurographics Association, Goslar, Germany (2012). DOI 10.2312/PE/NPAR/NPAR12/029-036
15. Girshick, A., Interrante, V., Haker, S., Lemoine, T.: Line Direction Matters: An Argument for the Use of Principal Directions in 3D Line Drawings. In: Proc. NPAR, pp. 43–52. ACM, New York (2000). DOI 10.1145/340916.340922
16. Gooch, A.A., Long, J., Ji, L., Estey, A., Gooch, B.S.: Viewing Progress in Non-Photorealistic Rendering Through Heinlein's Lens. In: Proc. NPAR, pp. 165–171. ACM, New York (2010). DOI 10.1145/1809939.1809959
17. Gooch, A.A., Willemsen, P.: Evaluating Space Perception in NPR Immersive Environments. In: Proc. NPAR, pp. 105–110. ACM, New York (2002). DOI 10.1145/508530.508549
18. Gooch, B.: Human Facial Illustrations: Creation and Evaluation using Behavioral Studies and Functional Magnetic Resonance Imaging. Ph.D. thesis, University of Utah, USA (2003)
19. Gooch, B., Reinhard, E., Gooch, A.A.: Human Facial Illustrations: Creation and Psychophysical Evaluation. ACM Transactions on Graphics **23**(1), 27–44 (2004). DOI 10.1145/966131.966133
20. Greenberg, S., Buxton, B.: Usability Evaluation Considered Harmful (Some of the Time). In: Proc. CHI, pp. 111–120. ACM, New York (2008). DOI 10.1145/1357054.1357074
21. Halper, N., Mellin, M., Herrmann, C.S., Linneweber, V., Strothotte, T.: Psychology and Non-Photorealistic Rendering: The Beginning of a Beautiful Relationship. In: Proc. Mensch & Computer, pp. 277–286. Teubner Verlag, Stuttgart, Leipzig, Wiesbaden (2003)
22. Halper, N., Mellin, M., Herrmann, C.S., Linneweber, V., Strothotte, T.: Towards an Understanding of the Psychology of Non-Photorealistic Rendering. In: Proc. Workshop Computational Visualistics, Media Informatics and Virtual Communities, pp. 67–78. Deutscher Universitäts-Verlag, Wiesbaden (2003)
23. Healey, C.G., Enns, J.T.: Perception and Painting: A Search for Effective, Engaging Visualizations. IEEE Computer Graphics and Applications **22**(2), 10–15 (2002). DOI 10.1109/38.988741
24. Healey, C.G., Tateosian, L., Enns, J.T., Remple, M.: Perceptually-Based Brush Strokes for Nonphotorealistic Visualization. ACM Transactions on Graphics **23**(1), 64–96 (2004). DOI 10.1145/966131.966135
25. Hertzmann, A.: A Survey of Stroke-Based Rendering. IEEE Computer Graphics and Applications **23**(4), 70–81 (2003). DOI 10.1109/MCG.2003.1210867
26. Hertzmann, A.: Non-Photorealistic Rendering and the Science of Art. In: Proc. NPAR, pp. 147–157. ACM, New York (2010). DOI 10.1145/1809939.1809957
27. Igarashi, T., Matsuoka, S., Tanaka, H.: Teddy: A Sketching Interface for 3D Freeform Design. In: Proc. SIGGRAPH, pp. 409–416. ACM, New York (1999). DOI 10.1145/311535.311602
28. Inglis, T.C., Kaplany, C.S.: Pixelating Vector Line Art. In: Proc. NPAR, pp. 21–28. Eurographics Association, Goslar, Germany (2012). DOI 10.2312/PE/NPAR/NPAR12/021-028
29. Isenberg, T., Freudenberg, B., Halper, N., Schlechtweg, S., Strothotte, T.: A Developer's Guide to Silhouette Algorithms for Polygonal Models. IEEE Computer Graphics and Applications **23**(4), 28–37 (2003). DOI 10.1109/MCG.2003.1210862
30. Isenberg, T., Neumann, P., Carpendale, S., Sousa, M.C., Jorge, J.A.: Non-Photorealistic Rendering in Context: An Observational Study. In: Proc. NPAR, pp. 115–126. ACM, New York (2006). DOI 10.1145/1124728.1124747
31. Jackson, C.D., Acevedo, D., Laidlaw, D.H., Drury, F., Vote, E., Keefe, D.: Designer-Critiqued Comparison of 2D Vector Visualization Methods: A Pilot Study. In: ACM SIGGRAPH Sketches & Applications. ACM, New York (2003). DOI 10.1145/965400.965505
32. Kaptein, M., Robertson, J.: Rethinking Statistical Analysis Methods for CHI. In: Proc. CHI, pp. 1105–1114. ACM, New York (2012). DOI 10.1145/2207676.2208557
33. Keefe, D.F., Karelitz, D.B., Vote, E.L., Laidlaw, D.H.: Artistic Collaboration in Designing VR Visualizations. IEEE Computer Graphics and Applications **25**(2), 18–23 (2005). DOI 10.1109/MCG.2005.34

34. Kerlinger, F.N., Lee, H.B.: Foundations of Behavioral Research, 4th edn. Wadsworth Publishing/Thomson Learning, London (2000)
35. Kim, S., Hagh-Shenas, H., Interrante, V.: Conveying Shape with Texture: An Experimental Investigation of the Impact of Texture Type on Shape Categorization Judgments. In: Proc. InfoVis, pp. 163–170. IEEE Computer Society, Los Alamitos (2003). DOI 10.1109/INFVIS.2003.1249022
36. Kim, S., Hagh-Shenas, H., Interrante, V.: Conveying Shape with Texture: Experimental Investigation of Texture's Effects on Shape Categorization Judgments. IEEE Transactions on Visualization and Computer Graphics **10**(4), 471–483 (2004). DOI 10.1109/TVCG.2004.5
37. Laidlaw, D., Kirby, R., Jackson, C., Davidson, J., Miller, T., da Silva, M., Warren, W., Tarr, M.: Comparing 2D Vector Field Visualization Methods: A User Study. IEEE Transactions on Visualization and Computer Graphics **11**(1), 59–70 (2005). DOI 10.1109/TVCG.2005.4
38. Laidlaw, D.H.: Loose, Artistic "Textures" for Visualization. IEEE Computer Graphics and Applications **21**(2), 6–9 (2001). DOI 10.1109/38.909009
39. Maciejewski, R., Isenberg, T., Andrews, W.M., Ebert, D.S., Sousa, M.C.: Aesthetics of Hand-Drawn vs. Computer-Generated Stippling. In: Proc. CAe, pp. 53–56. Eurographics Association, Goslar, Germany (2007). DOI 10.2312/COMPAESTH/COMPAESTH07/053-056
40. Maciejewski, R., Isenberg, T., Andrews, W.M., Ebert, D.S., Sousa, M.C., Chen, W.: Measuring Stipple Aesthetics in Hand-Drawn and Computer-Generated Images. IEEE Computer Graphics and Applications **28**(2), 62–74 (2008). DOI 10.1109/MCG.2008.35
41. Mandryk, R.L., Mould, D., Li, H.: Evaluation of Emotional Response to Non-Photorealistic Images. In: Proc. NPAR, pp. 7–16. ACM, New York (2011). DOI 10.1145/2024676.2024678
42. Martín, D., Arroyo, G., Luzón, M.V., Isenberg, T.: Example-Based Stippling using a Scale-Dependent Grayscale Process. In: Proc. NPAR, pp. 51–61. ACM, New York (2010). DOI 10.1145/1809939.1809946
43. Martín, D., Arroyo, G., Luzón, M.V., Isenberg, T.: Scale-Dependent and Example-Based Stippling. Computers & Graphics **35**(1), 160–174 (2011). DOI 10.1016/j.cag.2010.11.006
44. Mould, D., Mandryk, R.L., Li, H.: Emotional Response and Visual Attention to Non-Photorealistic Images. Computers & Graphics **36**(5), 658–672 (2012). DOI 10.1016/j.cag.2012.03.039
45. Rautek, P., Bruckner, S., Gröller, E., Viola, I.: Illustrative Visualization: New Technology or Useless Tautology? ACM SIGGRAPH Computer Graphics **42**(3), 4:1–4:8 (2008). DOI 10.1145/1408626.1408633
46. Saito, T., Takahashi, T.: Comprehensible Rendering of 3-D Shapes. ACM SIGGRAPH Computer Graphics **24**(3), 197–206 (1990). DOI 10.1145/97880.97901
47. Salesin, D.H.: Non-Photorealistic Animation & Rendering: 7 Grand Challenges. Keynote talk at NPAR (2002)
48. Santella, A., DeCarlo, D.: Visual Interest and NPR: an Evaluation and Manifesto. In: Proc. NPAR, pp. 71–78. ACM, New York (2004). DOI 10.1145/987657.987669
49. Schmidt, R., Isenberg, T., Jepp, P., Singh, K., Wyvill, B.: Sketching, Scaffolding, and Inking: A Visual History for Interactive 3D Modeling. In: Proc. NPAR, pp. 23–32. ACM, New York (2007). DOI 10.1145/1274871.1274875
50. Schumann, J., Strothotte, T., Raab, A., Laser, S.: Assessing the Effect of Non-Photorealistic Rendered Images in CAD. In: Proc. CHI, pp. 35–42. ACM, New York (1996). DOI 10.1145/238386.238398
51. Seifi, H., DiPaola, S., Enns, J.: Exploring the Effect of Color Palette in Painterly Rendered Character Sequences. In: Proc. CAe, pp. 89–97. Eurographics Association, Goslar, Germany (2012). DOI 10.2312/COMPAESTH/COMPAESTH12/089-097
52. Strothotte, T., Preim, B., Raab, A., Schumann, J., Forsey, D.R.: How to Render Frames and Influence People. Computer Graphics Forum **13**(3), 455–466 (1994). DOI 10.1111/1467-8659.1330455
53. Tietjen, C., Isenberg, T., Preim, B.: Combining Silhouettes, Shading, and Volume Rendering for Surgery Education and Planning. In: Proc. EuroVis, pp. 303–310. Eurographics Association, Goslar, Germany (2005). DOI 10.2312/VisSym/EuroVis05/303-310

54. Winnemöller, H., Feng, D., Gooch, B., Suzuki, S.: Using NPR to Evaluate Perceptual Shape Cues in Dynamic Environments. In: Proc. NPAR, pp. 85–92. ACM, New York (2007). DOI 10.1145/1274871.1274885
55. Winnemöller, H., Olsen, S.C., Gooch, B.: Real-Time Video Abstraction. ACM Transactions on Graphics **25**(3), 1221–1226 (2006). DOI 10.1145/1141911.1142018
56. Zhao, M., Zhu, S.C.: Sisley the Abstract Painter. In: Proc. NPAR, pp. 99–107. ACM, New York (2010). DOI 10.1145/1809939.1809951

15.A Data Resources

Some of the datasets used/created in the mentioned studies are available online for further analysis and future studies. For example, the following datasets are available online at the point of writing (of course, the URLs are always subject to change):

- sparse line drawing comparison by Cole et al. [6–8];
  → captured registered drawings, models, etc.:
  `http://gfx.cs.princeton.edu/proj/ld3d/`
- shape perception based on sparse line drawings by Cole et al. [6, 9];
  → gauge settings:
  `http://gfx.cs.princeton.edu/proj/ld3d/`
- evaluation of the pixelization of line art by Inglis and Kaplan [28];
  → user study data:
  `http://sites.google.com/site/tiffanycinglis/generating-pixel-art/generating-pixel-art---outlining`
- ethnographic study of illustrations by Isenberg et al. [30];
  → images:
  `http://www.cs.rug.nl/~isenberg/VideosAndDemos/Isenberg2006NPR`
- shape perception in dynamic 3D environments by Winnemöller et al. [54];
  → 3D models:
  `http://www.cs.northwestern.edu/~holger/Research/projects.htm`