Non-Photorealistic Animation and Rendering
Pierre Bénard and Holger Winnemoeller (Editors)

A Benchmark Image Set for Evaluating Stylization

David Mould[1†] and Paul L. Rosin[2‡]

[1]Carleton University, Ottawa, Canada
[2]Cardiff University, Cardiff, United Kingdom

Abstract

The non-photorealistic rendering community has had difficulty evaluating its research results. Other areas of computer graphics, and related disciplines such as computer vision, have made progress by comparing algorithms' performance on common datasets, or benchmarks*. We argue for the benefits of establishing a benchmark image set to which image stylization methods can be applied, simplifying the comparison of methods, and broadening the testing to which a given method is subjected. We propose a preliminary set of benchmark images, representing a range of possible subject matter and image features of interest to researchers, and we describe the policies, tradeoffs, and reasoning that led us to the particular images in the set.*

Categories and Subject Descriptors (according to ACM CCS): I.3.6 [Computer Graphics]: Methodology and Techniques—

1. Introduction

During the early days of a research topic, there is more focus on trailblazing than on formal analysis and evaluation. However, as the research area matures, many competing approaches are developed, and it becomes more difficult to distinguish between their relative benefits. In NPR, just as in other disciplines, a systematic and objective approach to comparative evaluation is necessary [Her10, Ise13].

An ideal method for evaluation should be general purpose, applicable to a wide variety of algorithms. The standard approach used in computer vision is to define a ground truth result against which an algorithm's results are compared. Unfortunately, for NPR no ground truth is available. Not only are many different stylizations possible (often radically different in appearance), but an individual stylization (e.g., etching) can come in many varieties. In computer vision, some "no-reference" image measures have been developed, which do not need ground truth images, and are generally based on low-level features extracted from the image. However, while this has proven popular for image quality assessment [MMB12], it is not easy to find "no-reference" measures for other assessment tasks. In addition, "no-reference" measures tend to lack discriminatory power compared to measures that have access to ground truth. While proxy measures [Her10] are fairly general and have been applied to NPR, they are at best loosely connected to the quantities of interest, such as the aesthetic appeal of the image.

Hall and Lehmann [HL13] agree with Hertzmann [Her10] in arguing that NPR cannot be assessed by human-subject experiments. Inspired by practices in Art History, they suggest that stylized images should be assessed by comparison to other existing (e.g. art) works, as well as existing criteria ("norms") used implicitly by people in the field, such as automation, algorithmic elegance, novelty, or "wow factor". This paper concentrates on facilitating comparison: the relative strengths and weaknesses of different algorithms can be revealed by applying them to a common dataset.

We use the term *benchmark* to refer to a standard set of data that algorithms use as input so as to produce comparable output. Usually, the evaluation is numerically scored, but that is not presently feasible in NPR. Nevertheless, an NPR benchmark can still provide a useful resource. At the most basic level, it facilitates comparison of NPR algorithms by providing a common set of images. Comparisons on common images already occur informally and sporadically, as images from some published papers are occasionally reused by later authors. Our intent is to encourage more systematic comparisons through use of a common dataset.

We propose an NPR benchmark, named *NPRgeneral*, in which the images collectively exhibit a wide range of possible features of interest, such as texture, contrast, complex edges, and semantically meaningful structures such as human faces. Details are given in Section 3. The benchmark can be used to compare algorithms, by inspecting the results of different algorithms on independently chosen input, and it can be used directly to help evaluate a single algorithm, showing the results over a variety of input images. Many of the images are quite challenging, and we do not expect every algorithm to succeed with every input. The failure cases are potentially of even more interest to the research community than the successes, since they embody unsolved problems and hence il-

† mould@scs.carleton.ca
‡ Paul.Rosin@cs.cf.ac.uk

DOI: 10.2312/exp.20161059

luminate directions for future work. This benchmark is not specific to any particular style or subject matter, and is intended for use by algorithms that can take arbitrary image input, hence the name "NPRgeneral".

2. Previous Work

Evaluation within the NPR discipline has been limited, both in terms of the amount of evaluation that has been carried out, and also regarding the variety of approaches taken to the evaluation [GPBOB08, Ise13]. Proxy metrics and variously formal and informal user studies are common. Mould [Mou14] argues for a principled form of subjective evaluation from proponents of stylization methods, to augment objective metrics and instead of user studies.

When the rendering style is tightly controlled, and moreover corresponds to a traditional artistic style, it is possible to obtain artists' drawings that can stand in for ground truth data. The similarity between artist and algorithmically generated images can then be compared by performing a user study. For example, Isenberg et al. [INC*06] compared a variety of pen-and-ink line drawing styles generated by human artists and algorithms. Images were shown to participants who were asked to sort the images into piles according to style, realism, aesthetics, or other considerations they thought helpful. While the participants could distinguish between the artist-generated and computer-generated drawings, the latter were still highly rated. The even more restricted task of drawing a single pencil line was explored by AlMeraj et al. [AWI*09]. Subjects were given the two-alternative forced choice task of deciding whether an image showed a line that was hand-drawn or computer-generated. Their tests indicated that the computer-generated line drawings were often perceived as hand-drawn.

An example of the proxy measure approach referred to by Hertzmann [Her10] is the memory game, used by Winnemöller et al. [WOG06] to evaluate their NPR algorithm [WOG06]. Participants were shown a 3×6 grid of cards with back side up; every time the player clicked a pair of cards they were revealed for a short time. If the cards uncovered by two consecutive clicks match, then both cards were removed; otherwise, they were turned back over to hide their contents. The time to complete the game and also the number of cards turned during the game were used to measure the performance of the player. When the memory game contained stylised images, the players' performances improved. From this, it was argued that the stylisation produced distinctive imagery. Other authors [GRG04, ZZ13, RL13] have also used matching tasks for evaluation of the authors' NPR algorithms, even though the purpose of the stylizations was not always or only to create memorable images.

Proper evaluation of image stylization methods requires comparisons between multiple approaches. Ideally, the algorithms would be run on common data so that meaningful conclusions could be drawn from the output; researchers should therefore coordinate on a common dataset. In computer graphics, informal reuse of well-known models is common, with models such as the bunny, angel, Buddha, and armadillo seen in many papers, and of course the ubiquitous teapot. Similarly, images such as Lena have seen informal and widespread usage in image processing papers. Stronger coordination becomes possible when researchers agree on a suitable benchmark dataset.

In recent years, image benchmarks have proliferated; there are now literally hundreds of publicly available benchmarks suitable for a wide range of topics, including analysis of faces, gestures, biometrics, object retrieval, pedestrian and vehicle tracking, medical images, character recognition, image segmentation, stereo, saliency, and more. For facial analysis alone, many such benchmark databases exist [BMP13]. Early efforts reused existing collections of photographs, such as Brodatz's *Photographic Album for Artists and Designers* [Bro66], which became popular for testing texture analysis algorithms. A later trend was to create bespoke image benchmarks, so as to enable careful control of the content. For example, the CMU PIE Database [SBB03] captured 41,368 face images of 68 people in 13 poses, with controlled lighting and facial expressions. Recently some extremely large benchmarks have been been created. For instance, the SUN Database [XHE*10] collected 130,519 images containing 99 categories from the Internet using online search queries for each scene category term, while the Large Scale Visual Recognition Challenge 2015 [RDS*15] used 150,000 images which had been collected from Flickr and other search engines, and then hand-labeled with the presence or absence of 1000 object categories. This year an image benchmark containing 3.4 million annotated images across 70 classes containing regions of interest has been released for Plankton Classification [OBPS15]. The largest image dataset of which we are aware is the YFCC100M dataset, containing one hundred million multimedia objects, 99M of which are photographs [TSF*16].

Thomee et al. [TSF*16] discuss some of the issues around image databases. While many image datasets have been proposed, most contain content with restrictive licenses, whether because the copyright owner must give permission for use, because the benchmark creator requires a license agreement as a condition of access, or because the benchmark is intended for use in a specific competition and access is restricted to competitors. The YFCC100M dataset contains only content with some sort of Creative Commons license. While Thomee et al. suggest that the massive size of YFCC100M is a key strength, its vastness poses problems as well. In NPR, where researchers labor over the evaluation of individual images, a small benchmark set is needed. Thomee et al. suggest mechanisms for communicating subset selection logic; in this paper, we directly propose a dataset of twenty images, small enough that all researchers should find it easy to apply their methods to all of them.

3. Principles of Image Dataset

Before presenting our set of selected images, we will discuss the policies that led us to that particular set. The set of images is by no means unique in satisfying our constraints; indeed, we expect the benchmark set to evolve quickly with input from the community. However, the set does serve to demonstrate the possibility of a plausible compromise among complex and sometimes opposing considerations.

Challenging images: The benchmark needs to include challenging images that are likely to be problematic for NPR algorithms. As different types of algorithms typically have different weaknesses,

the images should be challenging in different ways. This requirement will help uncover weak spots in current algorithms, and identify limitations which can be addressed in future work. Thus it helps enable a realistic appraisal of the state of the art (undercutting overselling algorithms) and will help push algorithmic development.

Range of difficulty: The benchmark should include images covering a range of levels of difficulty, so that the overall results of applying an NPR algorithm to the benchmark should not be a binary pass or fail. Indeed, a given image should not be a binary pass or fail, but some more complex measure of how effective the algorithm is, possibly qualitative. Including only challenging images would likely discourage many potential users, especially if they were developing more experimental methods.

Small number of images: Whereas the trend for benchmark datasets such as those used in computer vision is to contain thousands or millions of images, our requirement is the opposite. Image analysis methods that produce results such as classifications can easily compute assessments using automatic scoring. Conversely, most evaluation in NPR will be done manually. A small dataset is essential for manual evaluation to be manageable; given a large dataset, we expect that users would only use small selections, and since different users would make different selections, the results across different papers would not be comparable, defeating the original purpose of using a common benchmark. A sufficiently small dataset can be treated in its entirety.

Notice that that the criteria of using a small dataset and providing a broad coverage are in conflict. Since we cannot sacrifice the compactness of the dataset, its coverage is necessarily limited. In particular, semantic variation of the photographic subjects somewhat suffers. However, low-level details are still extremely varied within the set we chose.

Photographic images: The images in the dataset should be conventional photographs, as we think that stylizing captured real-world scenes offers the most difficult and widely relevant problems. There may be some utility in stylizing hand-drawn or other artistic images, but work (in sketch-based modeling, for example) that uses handmade images generally uses rich data including a history of marks and information about the primitives involved. We are not aware of work that concentrates on stylizing general handmade images using only the images themselves. Computer-generated synthetic images are another possibility; even more than handmade images, computer-generated images would typically contain much more than simply color information, with additional channels available such as depth, normal, object ID, and surface texture coordinate. These additional channels can potentially be exploited by stylization algorithms to good effect. We recommend creating an entirely separate dataset of 3D models and scenes for benchmarking evaluation of such methods.

Still images: We have deliberately excluded video from the present benchmark, not because we think it is unimportant, but rather because we think it is important enough to do a good job with it and it is distinct enough from images that its considerations will need to be addressed separately. For video, we have the added complications of time and motion. Complex motions and apparent motions owing to changes in camera parameters (focus, zoom, orientation) and the movement of the camera and of objects in the scene need to be considered carefully. Even basic questions like the appropriate duration of a shot do not have obviously correct answers.

Standard painting types: Many captured images follow standard topics. For instance, in the AVA dataset [MMP12] landscape, still life, animals, and portraits are all popular tags. NPR has been influenced by historical artistic practices, and standard painting genres such as landscapes, portraits, and still lifes should be well represented in the benchmark.

Aesthetics: Many stylization algorithms are designed with the intention of generating aesthetically pleasing results, and the trend for researchers in this area is to use source images that in their original state are also aesthetically pleasing.

Metadata: Including metadata such as numerical ratings of image characteristics is a useful adjunct, as these can then help characterise the performance of NPR algorithms. Correspondences between scores in the metadata and measures of image quality can be enlightening. For example, metadata could reveal that a specific algorithm has problems with images that are low contrast and contain large amounts of fine detail. The metadata can be provided by subjective human annotations and by objective measures using automatic image processing. We provided measurements of some characteristics of interest for the images in our proposed dataset.

Copyright clearance: Since NPR relies on manual evaluation (rather than listing numerical scores), it is essential that all the benchmark images have copyright clearance so that they can be published along with the derived results. We took images from Flickr, selecting only those whose license permits distribution of modified versions.

Image size: We wanted images for which large sizes – at least 2048 linear pixels – were available. We will make at least two sizes available for benchmark users: a large size and a smaller size, standardized at 1024 pixels width. Aspect ratios vary slightly; by chance, all our images had a landscape or square aspect ratio, but we did not particularly use aspect ratio as a selection criterion.

3.1. Image characteristics

The following is a list of image properties we sought to include. We selected images so that each property can be found in several images in the benchmark set. While not all properties are equally important, each property is doubtless of interest to some subset of stylization algorithms. For example, an algorithm may have an inherent scale parameter, and is is worthwhile to test how it copes with images where the elements vary in size. Many stroke-based algorithms have difficulty conveying fine-scale detail or high-frequency texture. Conversely, while filter-based algorithms with local thresholding can handle texture and fine detail well, long gradients may prove problematic. We do not intend to claim that the list is exhaustive; we welcome suggestions for additions that can guide the future development of the benchmark image set.

- **Variation in scale** of the elements in the image.
- **Fine detail**: high-frequency structure, whether fine-scale texture or semantically important elements that are quite small.

• **Variation in texture**, usually arising from multiple types and scales of texture within a single image.

• **Regular structure**, encompassing both regular patterns and clean shapes such as straight lines, 90-degree angles, and circular arcs.

• **Irregular texture** such as foliage or unkempt hair.

• **Visual clutter**: prominent visual elements that are irrelevant to communicating the main content of the image.

• **Vivid and varied colors** over the image.

• **Muted colors**, such that the image contains unsaturated colors and the color contrast is low.

• **Low contrast**: some important image elements have low dynamic range.

• **Mixed contrast**: different image regions have different dynamic ranges, or use similar dynamic ranges with different average intensities.

• **Complex edges**: some of the silhouettes or other important edges are long and complicated; the silhouette of a tree would be an example.

• **Thin features** such as wires or tree branches are present in the image.

• **Indistinct edges** where the semantics of the scene indicate an edge to a human observer, but the pixels exhibit only a small change in intensity or color.

• **Long gradients** of intensity or color in the image plane, perhaps due to curved surfaces or lighting changes.

• **Human faces** are of particular interest to human artists and audiences; we count only images where a face makes up a significant portion of the image, as in a portrait.

• **High key** (or generally light images) and **low key** (dark images) are included to confirm the robustness of the methods against more extreme inputs (which are nevertheless often generated for artistic effect).

3.2. Limitations

The principles articulated above provide guidelines for selecting images. However, these guidelines are not necessarily complete, and they leave considerable room for judgement in deciding precisely which images should be included. We do propose a specific set of images, discussed in the next section; we consider this to be "version 0.1" of the benchmark and we intend to update it with specific suggestions from the NPR community.

The principles constrain the benchmark content, sometimes with a negative impact on the applicability of the benchmark. By restricting our benchmark to a small set, we necessarily sacrifice detailed coverage of image variations. For machine learning applications, a much larger dataset is required. For specialized methods such as portraiture, many of our images are irrelevant and the benchmark is insufficient by itself.

The most salient constraint arises from our deliberate decision to exclude video from the current version of the benchmark. Image stylization methods can be applied to video straightforwardly, if not always effectively, by stylizing each video frame separately; a video benchmark set would help standardize evaluation of video stylization. As discussed above, though, video has many considerations that images lack. We concentrated on images in this paper, but intend to extend the benchmark to include video as well.

The benchmark also excludes items such as 3D scenes and models. While this to some extent reflects the research interests of the present paper's authors, we also think that the need for a benchmark set is not as crucial there, given de facto benchmarking in using common models such as the Stanford bunny.

This paper presents a basic version of the benchmark. Expanded benchmarks are possible. One vision of an expanded benchmark would be a hierarchical dataset, where a core subset would be considered mandatory, and then preselected sections of the full dataset could be used according to the requirements of the method. The risk of a larger dataset, even with a defined core, is that authors might be tempted to pick and choose subsets, undermining the usefulness of the common benchmark. Nonetheless, as the present general benchmark is unsuitable for specialized methods such as portraiture, we do envision at minimum a specialized "faces" module in a future benchmark; other special-purpose modules can be added if there is sufficient interest.

4. Proposed Benchmark Set

This section discusses our tentative benchmark set. All 20 images can be seen in Figure 1. Top row: dark woods; mountains; cabbage; Mac. Second: angel; barn; toque. Third: Oparara; arch; headlight. Fourth: Yemeni; daisy; snow. Fifth: athletes; desert; tomatoes. Last row: city; rim lighting; cat; berries.

Table 1 summarizes the list of image properties and shows which of our images possess them. The decision about whether or not to identify a given property with a given image is of necessity subjective, although our choices are informed by numerical measurements of related traits, summarized in Table 2. We measured *colorfulness, complexity, contrast, sharpness, lineness, noise,* and the mean and standard deviation of intensity. These low-level features vary widely over our image set, giving us some confidence that the benchmark provides a broad spectrum of test cases. In addition, they will enable ratings derived from user studies of NPR results to be correlated against image measures, so that relationships (e.g. a certain algorithm may perform poorly on noisy or low contrast images) can be easily identified. Details of the measurements are given next. Default parameters from the relevant papers are used unless otherwise stated.

Image complexity: computed following Machado and Cardoso [MC98], who encode the image using JPEG compression at a fixed quality factor quality; we used 50. For more complex images, compression will incur a high error, and also yield a low file size compression ratio. Therefore the ratio of these two terms is an estimate of the complexity of the original image.

Image colourfulness: computed following Hasler and Süsstrunk [HS03]. They use a simple measure involving the

	angel	arch	ath	barn	berr	cabb	cat	city	daisy	dark wood	desert	head light	mac	mtns	opa	rim	snow	toma	toque	yem
varied scale	✓	✓		✓		✓		✓	✓									✓		
fine detail		✓	✓	✓			✓				✓			✓	✓				✓	✓
varied texture		✓		✓	✓					✓	✓	✓		✓	✓	✓			✓	✓
regular				✓				✓				✓							✓	
irregular		✓		✓	✓		✓			✓	✓			✓	✓		✓			
clutter			✓		✓		✓								✓		✓	✓		
color		✓	✓	✓	✓													✓		
muted	✓					✓		✓		✓	✓	✓		✓						
low contrast	✓								✓		✓			✓			✓			
mixed contrast	✓					✓		✓		✓	✓	✓		✓	✓		✓			✓
thin features				✓		✓	✓	✓		✓	✓	✓	✓		✓		✓		✓	✓
complex edges			✓			✓	✓			✓	✓			✓		✓	✓	✓		✓
gradients		✓							✓		✓	✓	✓				✓	✓		
indistinct	✓					✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓		
low key	✓									✓					✓	✓				
high key									✓				✓	✓			✓			
portrait													✓			✓			✓	✓

Table 1: *Our assessment of which images possess which properties of interest.*

means and standard deviations of the image pixels in the red-green and yellow-blue channels of opponent colour space, with the weighting of these terms determined by a perceptual experiment.

Image contrast: computed following Matković et al. [MNN*05]. A non-linear mapping is applied to the image intensities to match them better to human perception. Each pixel's local contrast is measured as the mean absolute difference with respect to its neighbouring four pixels, and contrast is summed over the image. The process is repeated at multiple (specifically 9) image resolutions, and a weighted sum of these contrasts provides the final measure.

Image sharpness: computed following Bahrami and Kot [BK14], who compute for each pixel the maximum difference with respect to its 8-neighborhood, termed the maximum local variation MLV. The distribution of MLV in an image is modelled by a Generalized Gaussian Distribution (GGD) with a weighting to increase sensitivity to large MLV values. The sharpness measure is taken as the standard deviation of the GGD.

Image noise estimation: computed following Immerkaer [Imm96]. His approach assumes that the estimation should be insensitive to edges, and so the image is convolved with a Laplacian, which should give no response at edges. Assuming normally distributed noise, the noise level is derived from the Laplacian response. None of our images is particularly noisy; for our purposes the measurement is better thought of as an estimate of the image's high-frequency content.

Lineness: since a standard approach was not available, we developed a new measure. In a similar manner in which the summed edge strength over the image is used to measure sharpness [RRAJ15], we have used the absolute value of the Laplacian of Gaussian (LoG) at two scales ($\sigma = \{2,4\}$) to measure the response to dark or bright lines. However, the LoG also produces responses adjacent to edges, and so these have been suppressed following the approach taken by Rosin and Lai [RL13].

Of course, the image measures are not mutually independent, and the covariance matrix of the values in Table 2 reveals their relationships. Since the numerical ranges of the measures are not standardised we have first normalised each measure to have unit standard deviation.

$$\begin{pmatrix} \textit{color} & \textit{complex} & \textit{contrast} & \textit{sharp} & \textit{line} & \textit{mean} & \textit{stdev} & \textit{noise} \\ 1 & .315 & .372 & .039 & .174 & .197 & .427 & .314 \\ .315 & 1 & .623 & .559 & .787 & -.108 & .107 & .994 \\ .372 & .623 & 1 & .541 & .607 & -.378 & .605 & .615 \\ .039 & .559 & .541 & 1 & .698 & -.446 & .066 & .550 \\ .174 & .787 & .607 & .698 & 1 & .005 & .276 & .738 \\ .197 & -.108 & -.378 & -.446 & .005 & 1 & .177 & -.128 \\ .427 & .107 & .605 & .066 & .276 & .177 & 1 & .103 \\ .314 & .994 & .615 & .550 & .738 & -.128 & .103 & 1 \end{pmatrix}$$

It can be seen that there is a strong correlation between complexity and the noise measure. The lineness measure also has high correlations between complexity, noise, and sharpness. A set of lower, but still reasonably high, correlations exist between contrast and complexity, noise, lineness, and the standard deviation of intensity. Note that no significant correlations exist between any image measures and colourfulness or mean intensity. Thus we see that, despite some correlations, the image measures still capture a reasonable range of image characteristics.

In the remainder of this section, we discuss the individual images in our benchmark. Each image has a combination of low-level and higher-level features of interest. Not all features are of equal importance, nor equally widespread throughout the benchmark set; a simple count of features does not give a very good estimate of the value of a particular image. In addition to seeking variety of content and image features in the set, we tried to make all the individual images reasonably appealing.

Angel. The stone of this image is fairly dark overall, but high intensities along the angel's arm and torso produce areas of high con-

	colourfulness	complexity	contrast	sharpness	lineness	mean	standard deviation	noise
angel	12.97	0.48	4.84	0.19	3.32	67.77	**29.96**	3.48
arch	71.48	**2.98**	**10.68**	0.21	**8.20**	113.00	62.14	**24.05**
athletes	45.68	0.21	5.50	0.16	2.43	136.37	47.76	1.54
barn	73.39	0.82	6.08	0.18	4.59	142.15	63.10	6.08
berries	**101.74**	0.71	9.38	0.20	4.46	105.74	66.24	5.30
cabbage	26.18	0.39	5.40	0.16	4.26	108.64	38.30	2.55
cat	38.39	0.93	7.18	**0.22**	4.37	112.73	55.49	8.86
city	47.86	0.48	6.76	0.18	6.14	148.45	72.47	2.80
daisy	32.04	0.08	**1.99**	**0.08**	1.48	**208.87**	32.47	**0.59**
darkwoods	34.81	1.32	6.69	0.20	4.69	55.17	39.78	10.36
desert	46.59	0.52	4.29	0.13	2.57	114.06	39.54	4.88
headlight	24.49	0.15	6.28	0.14	4.02	93.86	59.71	0.88
mac	57.52	**0.07**	2.58	0.12	**1.09**	159.93	46.03	1.00
mountains	41.64	0.13	5.09	0.10	1.38	145.03	64.74	1.25
oparara	27.65	0.56	6.88	0.18	3.45	44.71	42.54	4.09
rim	**11.98**	0.21	5.62	0.21	2.26	**23.72**	42.34	2.51
snow	20.73	0.79	4.76	0.22	6.47	183.55	58.11	5.86
tomato	60.68	0.21	6.63	0.15	2.05	90.31	69.47	2.41
toque	23.10	0.36	9.14	0.14	3.11	121.71	**81.45**	3.00
yemeni	57.72	0.35	6.26	0.16	2.94	90.01	63.70	3.17

Table 2: *Numerical measurements of image properties for each of the images in the benchmark; minimum and maximum values are highlighted for each measure.*

trast. Lower contrast makes some important image elements difficult to see, such as the angel's nose and wing and the lower faces; overall, we assess the contrast as mixed. Elements exist over multiple scales, from the largest structures such as the arm and wings, to smaller structures such as facial features, feathers, and the leaves of the wreath. Colors are muted, and some edges are indiscernible owing to the lighting and low color contrast. Texture details in the stone surface add further visual interest.

Arch. This image depicts the Liberty Bell Arch in Nevada. It has strong and moderately interesting silhouettes, and it was included in the benchmark because of its irregular rock textures. There are high-frequency textures throughout the image, but the image-space scale of the rocks varies across the image, from the larger objects on the leftmost part to the smaller structures in the lower middle and right. The color range of the rocks is limited. The sky contributes a long vertical gradient. Communicating the sometimes indistinct structure of the plants and features of the rock will be a challenge for many stylization methods.

Athletes. Unlike the other images in the benchmark, we see the full human figure in this action scene. The high contrasts and bright colors make the image superficially straightforward for many methods, but there is potential for distraction from the irregular albeit blurry background, and some edges, such as the hair and the cleats, will be complex if the structure is preserved faithfully. Researchers will often want to preserve fine details of the facial expressions of the athletes; we do not label this image as a portrait, though, since the faces occupy so little of the image plane.

Barn. This colorful image contains objects over a wide range of scales, from the largest objects such as the barn and silo, through intermediate-sized objects such as trees and the component parts of the buildings, down to very small structures such as tree branches, boards on the barn's front, and the ladder leading up the silo. Many features are thin, including tree branches and the struts and rafters visible on the nearest part of the barn. Texture is varied, with irregular texture in the vegetation and more regular texture on the silo and the face of the barn.

Berries. This is the most colorful image in our collection, as judged both subjectively and by the automatic "colourfulness" measure. It contains objects of somewhat different sizes – the strawberries, blackberries, blueberries, and spoon – and one could consider the image to be cluttered; not only is the pattern on the plate a potential distraction, unusually, the image is an example of foreground clutter, where not all details in the foreground necessarily need to be retained in order to communicate the image content. There is a mix of edge strengths. The overall image might be considered a variable texture, and the textures on the strawberries and blackberries differ.

Cabbage. This image has little color range but a wide range of intensities. The leaf boundaries are convoluted and sometimes difficult to detect; in places, they can be confused by the interior edges of the leaf veins. The veins themselves are thin features and occur at different scales, being larger on the outer leaves than the inner ones. The lighting is varied over the image. We anticipate the cabbage being a moderately challenging image for stylization methods.

Cat. The complex patterns and detail in the fur of the cat provide most of the visual interest of this image. The blurry but varied background may be challenging, with an indistinct boundary separating it from the furry foreground. The cat's whiskers are thin but

definite features. Edge shapes in this image (e.g., the fur of the cat's ears) will be complex even when well defined.

City. The masterly composition of this image provides a high level of visual interest throughout. Colors are generally muted, but the contrast is usually high; the dark clothing of the human figures provides a focal point. In the city itself, windows and building silhouettes are regular structures, while more distant buildings have reduced contrast and ultimately vanish. Wiring in the interior and architectural elements on building exteriors are thin features. The perspective yields structures over a wide range of scales.

Daisy. A high-key image with some sharp and some blurry edges. The petals vary in size considerably; gradients across the largest petals, caused both by soft shadows and by curvature, offer a mild complication to algorithms. The central texture is quite regular; anisotropic textures along several petals provide a different regular texture. Most of the image has little contrast, as the dynamic range is low in the first place. The image would be more challenging if it were less abstract, but nonetheless provides a way to weakly test methods on a large number of possible image features.

Dark Woods. A generally low-key image, the majority of the content of this photograph is the complex, irregular textures from the tree trunks and foliage. The trees themselves supply thin features to test algorithms. Contrast is variable, with low contrast in some of the more shadowed areas and stronger contrast between the darker trees and sunlit leaves behind them.

Desert. A composition with mixed texture, structure, and smooth gradients along the sand dunes. The colors are muted but there is a variety of intensity edges, including simple edges such as the lighter sand against the shadows and the darker region behind, and more complex and indistinct edge shapes such as the low-contrast texture edges in the uppermost region and the tree branch silhouettes. The mix of content plus generally low contrast makes this a challenging image.

Headlight. An image with regular patterns of variable contrast. Long gradients across the curving metal offer challenges to segmentation methods and threshold-based techniques. Reflections on the paint, as well as the grille, contain indistinct edges. Although it contains a recognizable object, the regular geometry makes the image seem a little abstract as well.

Mac. A portrait of a Mac user, with generally light tones. The man's features are partially occluded by the Mac, slightly complicating stylization; the presence of glasses and facial hair may also pose a problem for some dedicated portraiture methods. Though the glasses are very clear, they are thin features; there is some small-scale texture across the man's forehead. Some edges are blurred owing to the shallow depth of field, and the Mac itself supplies large-scale gradients.

Mountains. An overall light image owing to the mix of snow and cloud. Snow on the mountaintops provides irregular texture. The contrast is overall low. Some edges, such as those within the clouds, or the blue mountain against blue sky, are indistinct, but the strong silhouette of the trees provides a definite and complex edge shape for stylization algorithms to work on.

Oparara. This depicts a limestone arch over the Oparara River in New Zealand. It is unusually dark for a photograph, but its dark areas contain some variation and texture. The textures in the image are highly varied, including multiple scales and types of rock surface, ripples on the river, and foliage seen through the arch. We considered the image slightly cluttered, as the details of the rainforest visible through the arch are probably unimportant, and the silhouette of the arch is obscured by hanging vegetation. This image is likely to prove a challenge to many stylisation algorithms.

Rim Lighting. A portrait with a clean background and strong rim lighting. The darkest image in the benchmark, this image can be used to test algorithms for failures on near-uniform backgrounds. The high contrast along the rim may mask weaker but important contrasts on the man's facial features and clothing. In general, though, we do not expect this image to be especially challenging; it is a basic sanity check.

Snow. This is a largely high-contrast image that nonetheless may be challenging because of its overall light tone and the weak contrasts of some snow-covered branches in the midground. Dense arrangements of branches form irregular textures, while more prominent branches are thin features. There is also some muted texture on the barn. The silhouettes of the treetops and the branches against the barn are complex edges. While this image might be difficult to convey thoroughly in a stylization, we expect that its straightforward semantics may make it reasonably forgiving.

Tomato. The still-life composition of a bowl of tomatoes contains many features of interest. It has good contrast and strong colors as well as fine details (the hairs on the turnip root, the texture on the table and the curtain). The image contains structure across multiple scales – fine-scale texture and small structures such as stems and the tiny flower, medium-scale tomatoes, and the bowl and curtain at the largest scale. The curtain might be considered clutter. Still, the clarity of the composition and the overall clean edges will probably make this one of the simpler images to treat with image stylization techniques.

Toque. This is a largely straightforward portrait image, whose subject shows well-defined facial features. The regular knitted textures of the toque and scarf offer some interest; the relatively fine texture of the toque, combined with the lighting gradient, is especially noteworthy. Smaller gradients across the jacket, showing its shape, may or may not be preserved through stylization. The background, although very blurred, has high contrast. Finally, some regions of the silhouette are fairly complex, such as the fuzzy detail of the toque, the hairs on the image left, and the fur on the lower right.

Yemeni. A portrait of a man from Yemen, his strong features providing some inherent interest while including complications such as deep lines and a variable beard. The texture and coloration of his headgear afford additional opportunities for stylization. The shadows and lighting provide a challenge; some strong intensity edges, such as those on the tip of the man's nose, are unimportant, while weaker edges such as those on the right half of the man's face are critical.

5. On Adoption of the Benchmark

The benchmark is only of any benefit if the NPR community actually uses it. We envision two main use cases. First, researchers can include selected benchmark results in the pages of their published papers. Since the full benchmark has been kept sufficiently compact to be displayed in a single page, it is also feasible for them to include full results within their papers. Second, researchers can provide a more extensive set of benchmark results on a project page, augmenting the publication and helping future researchers by making comparisons easy.

We believe that widespread adoption of the benchmark will benefit the science of non-photorealistic rendering. It will encourage a more systematic approach to evaluation and a thorough disclosure of algorithms' behaviour so that weaknesses can be known and addressed by followup work. It will also help researchers in other ways, by providing sensible defaults for testing image stylization algorithms and improving access to past results for purposes of comparison.

At the same time, we recognize that benchmarks have drawbacks. A benchmark that is not representative of real data will lead to conclusions of dubious validity; we have tried to make our dataset as broad as possible, and anyway do not expect that researchers will concentrate single-mindedly on the benchmark to the exclusion of other images. The related issue of overfitting is a serious potential problem, where algorithms are finely tuned to the benchmark data and do not attain equally good performance on other data. In fact, the problem is worse in the absence of a benchmark, since researchers are free to choose inputs where their algorithms perform well; an independently chosen dataset eliminates the suspicion that the inputs were excessively curated. Lastly, saturation is a potential long-term issue, where the benchmark was at first challenging but the discipline has advanced to the point that these images are simple. In the absence of quantitative evaluation, and given the diversity of possible objectives for stylization algorithms, we do not think that saturation will be a problem in the non-photorealistic rendering field.

We encourage researchers to apply their algorithms to the entire dataset, not only to a subset. Some benefits, such as the transparency of using a dataset chosen independently rather than by the researchers, only become possible when the entire dataset is used. In cases where the algorithm is only meant to apply to a certain image type – e.g., specialized methods for portrait rendering – the appropriate subset of the benchmark can be extracted. Where algorithms are intended for more general use, however, the entire benchmark set should be shown; even if only selected images will fit into the paper, the benchmark results can be reported as supplementary material.

The dataset in this paper is tentative, and should be considered "version 0.1." As we get suggestions from the community, we will make necessary changes and fill any discovered gaps in the list of image properties, before finalizing the benchmark for a target date of August 2016. Of course, we hope that changes can be minimal, since we have striven to prepare a comprehensive dataset in the first place. Once version 1.0 is available, researchers can begin using it to test their algorithms.

6. Conclusion and Future Work

We presented NPRgeneral, a set of benchmark images to test image stylization algorithms; more importantly, we articulated a set of considerations that can guide the development of future benchmark sets. The image set should be large enough to include all the features of interest, but not so large that it becomes unwieldy for manual assessment. Features of interest include low-level features such as variable contrast and high-frequency structure, as well as high-level features such as human faces and clutter. Pragmatically, the images should be of adequate resolution and must be free of copyright encumbrances that would prevent distribution of modified images.

As the benchmark will only be beneficial when researchers use it, this paper is also a call to action to the community to take up the benchmark and report the results of your new and old algorithms. Researchers can showcase benchmark results in their papers as well as hosting the results of the full benchmark on a project page.

The set of images presented in this paper are "version 0.1" of the benchmark, and a 1.0 release will be established by August 2016, pending modifications and additions suggested by the community. The benchmark image data will be publicly available from `expressive.graphics/benchmark` as well as from `gigl.scs.carleton.ca/benchmark`.

Future versions of the benchmark may be extended to include video, or other image types such as depth images or plenoptic images. Often, stylization methods build on standard methods; we can facilitate comparisons by providing salience maps and presegmented images, for example, and perhaps other forms of standard preprocessing would be helpful. Additional metadata in the form of manual annotations of the images – e.g., manual foreground/background segmentation, or labelings of regions of interest – could be included in later versions of the benchmark.

Acknowledgements

All original images came from Flickr, provided by these photographers: Eole Wind (angel), Nathan Congleton (athletes), MrClean1982 (barn), HelmutZen (berries), Leonard Chien (cabbage), Theen Moy (cat), Rob Schneider (city), mgaloseau (daisy), JB Banks (dark woods), Charles Roffey (desert), Photos By Clark (headlight), Martin Kenny (Mac), jjjj56cp (mountains), trevorklatko (Oparara), James Marvin Phelps (arch), Paul Stevenson (rim lighting), John Anes (snow), Greg Myers (tomatoes), sicknotepix (toque), Richard Messenger (Yemeni).

References

[AWI*09] ALMERAJ Z., WYVILL B., ISENBERG T., GOOCH A. A., GUY R.: Automatically mimicking unique hand-drawn pencil lines. *Computers & Graphics 33*, 4 (2009), 496–508. 2

[BK14] BAHRAMI K., KOT A. C.: A fast approach for no-reference image sharpness assessment based on maximum local variation. *IEEE Signal Processing Letters 21*, 6 (2014), 751–755. 5

[BMP13] BOUSMALIS K., MEHU M., PANTIC M.: Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools. *Image Vision Computing 31*, 2 (2013), 203–221. 2

[Bro66] BRODATZ P.: *Textures : a photographic album for artists and designers*. Dover Publications, New York, 1966. 2

[GPBOB08] GATZIDIS C., PAPAKONSTANTINOU S., BRUJIC-OKRETIC V., BAKER S.: Recent advances in the user evaluation methods and studies of non-photorealisitic visualisation and rendering techniques. In *Proc. Info. Vis.* (2008), pp. 475–480. 2

[GRG04] GOOCH B., REINHARD E., GOOCH A.: Human facial illustrations: Creation and psychophysical evaluation. *ACM Trans. Graph. 23*, 1 (2004), 27–44. 2

[Her10] HERTZMANN A.: Non-photorealistic rendering and the science of art. In *Proceedings of the 8th International Symposium on Non-Photorealistic Animation and Rendering* (2010), pp. 147–157. 1, 2

[HL13] HALL P., LEHMANN A.-S.: Don't measure – appreciate! NPR seen through the prism of art history. In *Image and Video-Based Artistic Stylisation*, Rosin P. L., Collomosse J. P., (Eds.). Springer, 2013, pp. 333–351. 1

[HS03] HASLER D., SÜSSTRUNK S.: Measuring colourfulness in natural images. In *Proc. SPIE Human Vision and Electronic Imaging* (2003), pp. 87–95. 4

[Imm96] IMMERKAER J.: Fast noise variance estimation. *Computer Vision and Image Understanding 64* (1996), 300–302. 5

[INC*06] ISENBERG T., NEUMANN P., CARPENDALE S., SOUSA M. C., JORGE J. A.: Non-photorealistic rendering in context: an observational study. In *ACM Symp. NPAR* (2006), pp. 115–126. 2

[Ise13] ISENBERG T.: Evaluating and validating non-photorealistic and illustrative rendering. In *Image and Video-Based Artistic Stylisation*, Rosin P. L., Collomosse J. P., (Eds.). Springer, 2013, pp. 311–331. 1, 2

[MC98] MACHADO P., CARDOSO A.: Computing aesthetics. In *Brazilian Symposium on Artificial Intelligence* (1998). 4

[MMB12] MITTAL A., MOORTHY A. K., BOVIK A. C.: No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing 21*, 12 (2012), 4695–4708. 1

[MMP12] MURRAY N., MARCHESOTTI L., PERRONNIN F.: AVA: A large-scale database for aesthetic visual analysis. In *Conf. Computer Vision and Pattern Recognition* (2012), pp. 2408–2415. 3

[MNN*05] MATKOVIĆ K., NEUMANN L., NEUMANN A., PSIK T., PURGATHOFER W.: Global contrast factor – a new approach to image contrast. In *Computational Aesthetics* (2005), pp. 159–167. 5

[Mou14] MOULD D.: Authorial subjective evaluation of non-photorealistic images. In *Proceedings of the Workshop on Non-Photorealistic Animation and Rendering* (New York, NY, USA, 2014), NPAR '14, ACM, pp. 49–56. URL: http://doi.acm.org/10.1145/2630397.2630400, doi:10.1145/2630397.2630400. 2

[OBPS15] ORENSTEIN E. C., BEIJBOM O., PEACOCK E. E., SOSIK H. M.: WHOI-Plankton – A large scale fine grained visual recognition benchmark dataset for plankton classification. *CoRR abs/1510.00745* (2015). 2

[RDS*15] RUSSAKOVSKY O., DENG J., SU H., KRAUSE J., SATHEESH S., MA S., HUANG Z., KARPATHY A., KHOSLA A., BERNSTEIN M., BERG A. C., FEI-FEI L.: ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision 115*, 3 (2015), 211–252. 2

[RL13] ROSIN P. L., LAI Y.-K.: Artistic minimal rendering with lines and blocks. *Graphical Models 75*, 4 (2013), 208–229. 2, 5

[RRAJ15] REDI M., RASIWASIA N., AGGARWAL G., JAIMES A.: The beauty of capturing faces: Rating the quality of digital portraits. In *Conf. on Automatic Face and Gesture Recognition* (2015), pp. 1–8. 5

[SBB03] SIM T., BAKER S., BSAT M.: The CMU pose, illumination, and expression database. *IEEE Trans. Pattern Anal. Mach. Intell. 25*, 12 (2003), 1615–1618. 2

[TSF*16] THOMEE B., SHAMMA D. A., FRIEDLAND G., ELIZALDE B., NI K., POLAND D., BORTH D., LI L.-J.: YFCC100M: The new data in multimedia research. *Commun. ACM 59*, 2 (Jan. 2016), 64–73. URL: http://doi.acm.org/10.1145/2812802, doi:10.1145/2812802. 2

[WOG06] WINNEMÖLLER H., OLSEN S., GOOCH B.: Real-time video abstraction. *ACM Trans. Graphics 25*, 3 (2006), 1221–1226. 2

[XHE*10] XIAO J., HAYS J., EHINGER K. A., OLIVA A., TORRALBA A.: SUN database: Large-scale scene recognition from abbey to zoo. In *Conference on Computer Vision and Pattern Recognition* (2010), pp. 3485–3492. 2

[ZZ13] ZHAO M., ZHU S.-C.: Abstract painting with interactive control of perceptual entropy. *ACM Transactions on Applied Perception (TAP) 10*, 1 (2013), 5. 2

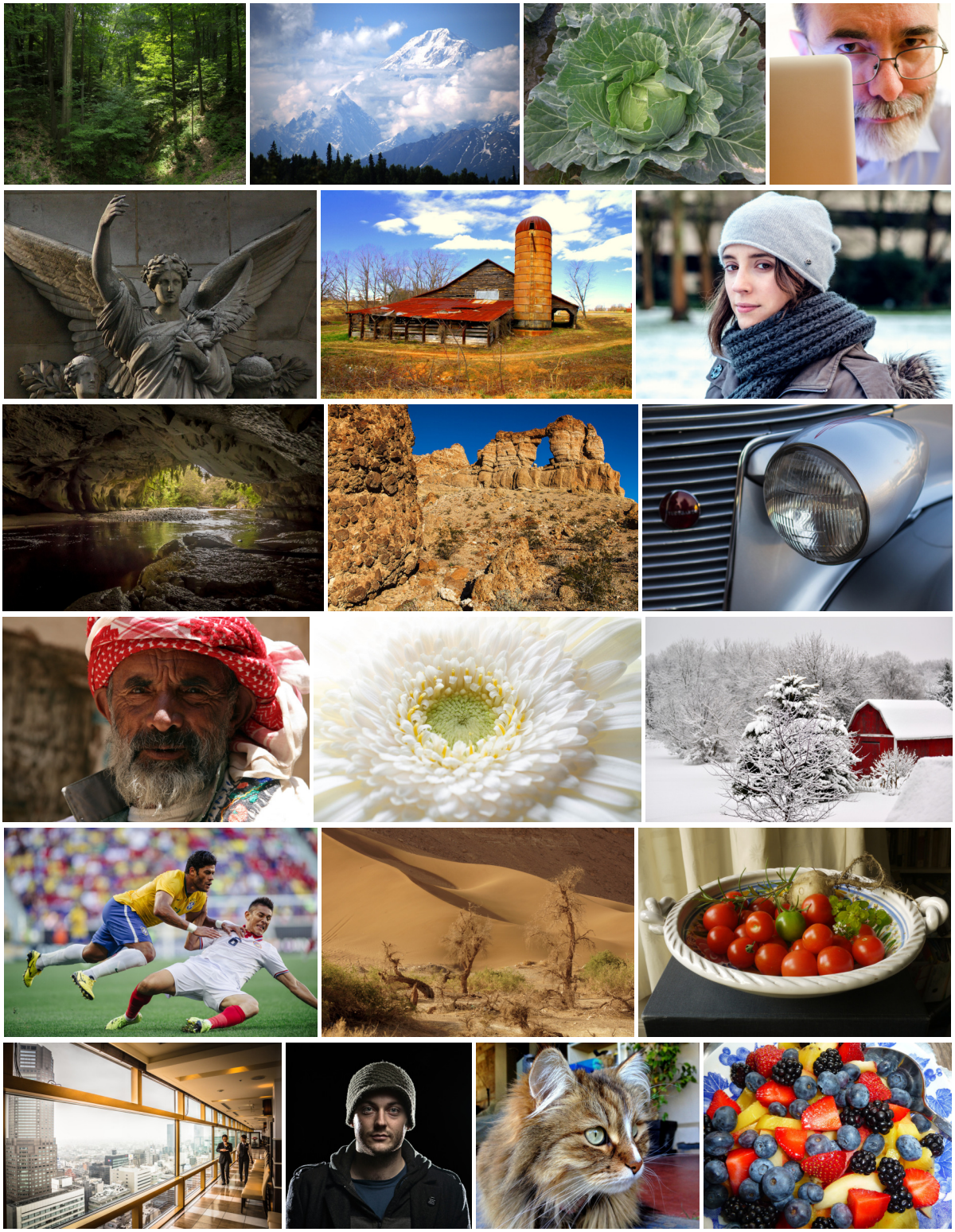

Figure 1: *The set of 20 benchmark images.*