Technical Section

Developing and applying a benchmark for evaluating image stylization

David Mould\textsuperscript{a}\textsuperscript{*}, Paul L. Rosin\textsuperscript{b}

\textsuperscript{a} Carleton University, Canada
\textsuperscript{b} Cardiff University, United Kingdom

\textbf{A B S T R A C T}

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 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. Then, we apply six previously existing stylization algorithms to the benchmark images; we discuss observations arising from the interactions between the algorithms and the benchmark images. Inasmuch as the benchmark images were able to thoroughly exercise the algorithms and produce new material for discussion, we can conclude that the benchmark will be effective for its stated aim.

© 2017 Elsevier Ltd. All rights reserved.

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 [1–3].

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 (etching, say) 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 [4], 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 [1] 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 [5] agree with Hertzmann [1] 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 \textit{NPGeneral}, 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 illuminate 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”.

This paper is an extended version of the conference paper that initially introduced the benchmark set [6]. In the current paper, we recapitulate the discussion of the need for an NPR benchmark and the reasoning behind the specific benchmark images chosen; our new contribution is to apply six existing stylization algorithms to the full benchmark set and assess the results. The discussion of the stylized benchmark images serves both as an example of how we imagine others using the benchmark in their own future papers, and as a demonstration of the effectiveness of the benchmark set: the benchmark contains sufficient variety of content that we can gain some insight into the behavior of stylization algorithms by examining the stylized benchmark images.

Note that our goal in this paper is to present and assess the benchmark for its ability to exercise image stylization methods. We use some existing methods as examples to demonstrate the breadth of content in the benchmark image set. Turning a critical eye to the filtered images, we will point out particular aspects that strike us as noteworthy, as arising from interactions between the stylization algorithms and the contents of the benchmark images.

We are not, per se, making a general assessment of the effectiveness of the existing methods, nor making direct comparisons between the methods we discuss. The reader can consult the original papers to see the objectives of the original authors and their evaluation of the methods’ effectiveness.

With dozens or perhaps hundreds of image stylization algorithms available in the literature, we must be selective in this paper. We chose six stylization methods, grouped into two broad categories: abstraction, in the sense of stylization through detail removal, and reduced-palette rendering, where the image is communicated without color and using only a limited tonal range. Both categories represent overall objectives shared by numerous methods in NPR. Within each group, we chose three recent algorithms to apply. Discussion of the benchmark images themselves, and our observations on the interaction between the chosen stylization methods and the benchmark images, make up the bulk of this paper.

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 [2,7]. Proxy metrics and variously formal and informal user studies are common. Mould [3] 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. [8] 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. [9]. 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 [1] is the memory game, used by Winnemöller et al. [10] to evaluate their NPR algorithm [10]. 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 stylized images, the players’ performances improved. From this, it was argued that the stylization produced distinctive imagery. Other authors [11–13] 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, 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 [14]. Early efforts reused existing collections of photographs, such as Brodatz’s Photographic Album for Artists and Designers [15], 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 [16] captured 41,368 face images of 68 people in 13 poses, with controlled lighting and facial expressions. Recently some extremely large benchmarks have been created. For instance, the SUN Database [17] 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 [18] 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. Recently an image benchmark containing 3.4 million annotated images across 70 classes containing regions of interest was released for Plankton Classification [19]. The largest image dataset of which we are aware is the YFCC100M dataset, containing one hundred million multimedia objects, 99M of which are photographs [20].

Thomee et al. [20] 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 bench-
دریافت فوری متن کامل مقاله

امکان دانلود نسخه تمام متن مقالات انگلیسی
امکان دانلود نسخه ترجمه شده مقالات
پذیرش سفارش ترجمه تخصصی
امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
امکان دانلود رایگان ۲ صفحه اول هر مقاله
امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
دانلود فوری مقاله پس از پرداخت آنلاین
پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات