Benchmarking Non-Photorealistic Rendering of Portraits

Paul L. Rosin*
Cardiff University
UK

David Mould†
Carleton University
Canada

Itamar Berger
The Interdisciplinary Center Herzliya
Israel

John Collomosse
University of Surrey
UK

Yu-Kun Lai
Cardiff University
UK

Chuan Li
Lambda Labs, Inc.
USA

Hua Li
Canada

Ariel Shamir
The Interdisciplinary Center Herzliya
Israel

Michael Wand
Mainz University
Germany

Tinghuai Wang
Nokia Labs, Nokia Technologies
Finland

Holger Winnemöller
Adobe Systems, Inc.
USA

ABSTRACT

We present a set of images for helping NPR practitioners evaluate their image-based portrait stylisation algorithms. Using a standard set both facilitates comparisons with other methods and helps ensure that presented results are representative. We give two levels of difficulty, each consisting of 20 images selected systematically so as to provide good coverage of several possible portrait characteristics. We applied three existing portrait-specific stylisation algorithms, two general-purpose stylisation algorithms, and one general learning based stylisation algorithm to the first level of the benchmark, corresponding to the type of constrained images that have often been used in portrait-specific work. We found that the existing methods are generally effective on this new image set, demonstrating that level one of the benchmark is tractable; challenges remain at level two. Results revealed several advantages conferred by portrait-specific algorithms over general-purpose algorithms: portrait-specific algorithms can use domain-specific information to preserve key details such as eyes and to eliminate extraneous details, and they have more scope for semantically meaningful abstraction due to the underlying face model. Finally, we provide some thoughts on systematically extending the benchmark to higher levels of difficulty.

CCS CONCEPTS

•Computing methodologies → Non-photorealistic rendering; Image processing;

*corresponding author, email: Paul.Rosin@cs.cf.ac.uk
†corresponding author, email: Mould@scs.carleton.ca

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
NPAR'17, Los Angeles, CA, USA
© 2017 ACM. 978-1-4503-5081-5/17/07…$15.00
DOI: 10.1145/3092919.3092921

KEYWORDS

evaluation, non-photorealistic rendering, image stylisation, portraits

ACM Reference format:
Paul L. Rosin, David Mould,, Itamar Berger, John Collomosse, Yu-Kun Lai, Chuan Li, Hua Li, Ariel Shamir, Michael Wand, Tinghuai Wang, and Holger Winnemöller. 2017. Benchmarking Non-Photorealistic Rendering of Portraits. In *Proceedings of NPAR'17, Los Angeles, CA, USA, July 28-29, 2017,* 12 pages.
DOI: 10.1145/3092919.3092921

1 INTRODUCTION

Progress in science is best served when there are means to evaluate and compare theories and methods. Quantitative analysis is most desirable, as this makes systematic evaluation objective, and may also enable it to be scaled up to large numbers of tests with minimal effort. Unfortunately, in some instances, appropriate objective analysis cannot be achieved, and so researchers have to fall back on subjective evaluation. In large part, this is the situation for non-photorealistic rendering (NPR), for which it is hard to compute a score that reflects the aesthetic qualities of the rendered output. Consequently, performance evaluation and the comparison of algorithms has not been pursued within the NPR community to the same degree as in computer vision.

The limited ability to evaluate objectively has perhaps led to the lack of standard benchmarks for evaluating NPR algorithms. Typically, authors use their own images, and sometimes reuse a few images from previous NPR papers. Recently, Mould and Rosin [2016] released the first NPR benchmark data set, named *NPRgeneral*, comprising 20 images selected to satisfy a number of criteria, and to cover a range of characteristics. However, it does not have much coverage of portraits. For algorithms specifically oriented towards stylising faces, a more focussed benchmark is needed.

This paper presents a benchmark image set specifically for portrait stylisation[1]. It is systematically designed to provide a range of the characteristics present in portraits, and to control the level of difficulty: it is organised into multiple levels, with each level more challenging than the one below. We present the first two levels in this paper and show stylisations of level 1 from six existing algorithms.

1.1 Portraiture in Art

Historically, portraits served many purposes. Universally, they were intended to memorialise some aspect of the person depicted. However, even from the earliest days, artists were not satisfied with producing an exact likeness of the subject. Rather, the portrait strayed from reproduction, possibly merely to flatter the subject, or possibly so as to convey some essential nature of the individual portrayed. Sometimes, the subject was someone of political or economic stature, and part of the mission of the portrait would be to communicate riches or grandeur. This could be achieved by surrounding the subject with symbolic elements, such as weapons or gold; other times, allusion would play a role, as in Bronzino's depiction of the Genoese admiral Andrea Doria in the guise of Neptune. A modern equivalent would be an illustrator depicting George Lucas as a Jedi knight. The range of poses and content in historical portraiture is broad; Figure 1 gives a few examples.

Figure 1: Painted portraits: a Maori chieftain by Lindauer; a portrait of a family by Rubens; Andrea Doria as Neptune by Bronzino; *Portrait of Maria Teresa de Vallabriga on horseback* by Goya; *Christina's World* by Wyeth. All images came from Wikimedia commons.

1.2 Portraiture in NPR

There is a long history of creating portraits in art: from painting and sculpture through to photography. Likewise, portraits have also figured in NPR, from the early days, such as Brennan's [Brennan 2007] work on computer-assisted caricature generation, up to recent trends such as the convolutional neural network approach [Selim et al. 2016]. Even those papers that do not specifically target faces often provide examples of portraits in the teaser or first figure [Galea et al. 2016; Haeberli 1990; Li and Wand 2016; Olsen and Gooch 2011; Winnemöller et al. 2012].

While general algorithms can be applied to images of faces, specialised algorithms have also appeared. In general, some knowledge of the domain can improve stylisation algorithms: methods can benefit from specialised models. Faces are an unusually important element of images, hence have received focussed attention [Berger et al. 2013; Colton et al. 2008; Wang et al. 2013; Zhao and Zhu 2013]. The restricted domain makes models feasible.

The use of an underlying model helps fix simple problems to which human viewers are especially sensitive. Small changes, such as the omission of an eye region, can cause the audience to perceive an image dramatically differently; even very subtle changes, such as the eye's pupil being moved, can influence audience perception.

1.3 Goals of this Paper

We have two main objectives in this work. First and foremost, we hope to help systematise evaluation of portrait stylisation. Researchers can show the results of applying their stylisation algorithms to the images in the benchmark, thus exercising the algorithms over a broad range of possible faces. Having a common set of faces will facilitate comparisons between different methods. When researchers show results from all benchmark images, readers will know that results have not been specially selected to favor the cases where the algorithm works well, or to disfavour cases where the algorithm fails.

A secondary objective for this paper is to stimulate further portrait research. While current face-specific methods are effective in the constrained situations for which they were designed, historical artworks show a vast range of face types, poses, and complications such that existing automated algorithms cannot cope. Similarly, even though many photographs use conventional poses, many do not; the depiction of people in photographs is enormously varied. We urge the community to investigate more robust algorithms that can deal with a broader range of input images.

We make three main contributions in this paper. First, we provide a roadmap for a multi-stage image benchmark for portrait stylisation, where the first level contains highly constrained images of the sort now used in portrait stylisation, and later levels introduce successively more difficult and more pronounced complications. Second, we provide two sets of 20 images each for the first two levels of the benchmark, and describe the detailed design process that led to these image sets. Third, we apply several stylisation algorithms, both general and face-specific, to the first level of the benchmark and discuss our findings. In brief, we found that the portrait-specific algorithms gain some robustness from the domain information, but performance degrades when the input images do not adhere to the constraints assumed by the face model.

2 PREVIOUS WORK

Mould and Rosin [2016] created an NPR benchmark data set, *NPRgeneral*, containing 20 images selected to include a variety of attributes and content, such as fine detail, long gradients, mixed contrast, and

[1] The benchmark image set can be downloaded from http://users.cs.cf.ac.uk/Paul.Rosin/NPRportrait as well as from http://gigl.scs.carleton.ca/benchmark.

human faces. As only one of nearly 20 possible elements, human faces were present in only a few images. However, because of the interest in human faces specifically, the existing benchmark does not suffice: the community would benefit from access to a different benchmark set that concentrates on faces.

Although NPR benchmarking is so limited, literally hundreds of publicly available benchmark data sets exist in the field of computer vision. From the early days, test images included portrait images such as Lena, Barbara, and Elaine. Subsequently, many data sets specifically consisting of faces were developed for benchmarking face detection and recognition algorithms. While the older data sets contain images of a few tens of subjects [Samaria and Harter 1994] the increasing focus on performance evaluation quickly led to larger data sets containing hundreds of subjects [Phillips et al. 2000]. Recently, large-scale data-driven machine learning has required massive data sets for training, such as the Facebook dataset [Becker and Ortiz 2013; Taigman et al. 2014].

In order to construct an NPR portrait benchmark data set, images could be sourced from one or several existing computer vision data sets. However, the older data sets were collected with overly restrictive conditions. For instance, JAFFE [Lyons et al. 1999] contains only Japanese females. ORL [Samaria and Harter 1994] contains predominantly Caucasian males, the faces are very tightly cropped, and the images are low resolution. Later, larger-scale efforts, such as DARPA's FERET programme [Phillips et al. 2000] are also not suitable: like the ORL and JAFFE, the image capture was too standardised, using the same physical setup and location (e.g. background) during construction of the data set. Moreover, grayscale rather than colour images were captured. In contrast, more recent data sets such as Labeled Faces in the Wild [Huang et al. 2007] are too unconstrained: they contain substantial variations in pose, background, lighting, and occlusion. In an effort parallel to ours but in the opposite direction, researchers at the University of Bath developed an image set with a range of depictions of people in artwork [Westlake et al. 2016]; since these images are not photographs, they are not suitable for conventional stylisation efforts.

The second aspect of performance evaluation is the need to define protocols to carry out the evaluation, and this topic has been considered within computer vision at great length. There is a large literature; work by Haralick [1994], Forstner [1996], and Thacker et al. [2008] is representative. In addition, many specific approaches to evaluation have been developed for individual tasks such as object recognition, edge detection, character recognition, line detection, etc. Unfortunately, evaluation of NPR is more problematic than for computer vision, as it is difficult to quantify the quality of a rendered image. Not only do standard image comparison measures such as PSNR or SSIM fail to capture important perceptual and aesthetic aspects of a stylised image, but NPR lacks a ground truth against which to perform comparison. Practitioners in NPR have considered these issues [Isenberg 2013], and have suggested to employ proxy measures [Hertzmann 2010] in place of directly evaluating the aesthetics of the stylised image, to carry out an authorial subjective evaluation [Mould 2014], or to compare the stylised image to art works and "norms" such as automation, algorithmic elegance, novelty, or "wow factor" [Hall and Lehmann 2013]; none of these is totally satisfactory. Another common approach is to perform a user study, although eliciting reliable user ratings for aesthetic judgements is not trivial [Mould 2014], and developing appropriate models of aesthetic judgement of artworks remains an ongoing topic of research [Leder et al. 2004; Palmer et al. 2013].

3 PRINCIPLES OF IMAGE DATASET

Many current NPR portrait systems are restricted to front-on single faces with simple backgrounds and no facial occlusion. We want to ensure that the benchmark is widely applicable; if it is too difficult then it will not be used by the community. At the same time, we would like to represent a fuller spectrum of possible images; in practice, people take a vast range of photographs, and it would be good to introduce some of these complications so as to push the capabilities of algorithmic image stylisation.

Hence, we plan to organise the benchmark into multiple stages, where the levels become increasingly unconstrained and more challenging. Within each level, we intend to produce a cross-section of possible complications. However, the difficulty level should not rise too quickly, or else progress may not be visible: a gradual increase in difficulty means that algorithms with small differences in robustness will have noticeably different ability to successfully process successive levels. The first level should be attainable by existing methods.

Our principles can be summarised as follows:

- The image set should contain a range of different face types. Furthermore, the images should present a broad collection of complications, capturing the range of conditions and environments where people photograph faces.
- The image set should be small enough that evaluating it manually is feasible.
- As a consequence of the tension between the first two principles, the benchmark should be organised into multiple levels of difficulty.
- The first level should correspond to the sorts of photographs used by existing portrait stylisation methods.
- The gap in difficulty between level n and level $n + 1$ should not be too great.

Our first level is governed by typical existing practices in portrait stylisation: adult faces in strict frontal views, neutral expression, clean backgrounds, with no ornamentation or facial hair. The second level relaxes these constraints slightly, permitting facial hair, mild facial expressions, a bit of jewellery, and more varied backgrounds. Levels 3 and 4 will relax constraints on background clutter, lighting, expressions, poses, and especially age range: they should include children and the elderly as well as adult faces.

For higher levels, complications abound. Even with straightforward poses and expressions, faces vary considerably. Textures – whether arising from wrinkles, scars, blemishes, or facial hair – are a potential source of difficulty, yet can be revealing and hence cannot be neglected. Hairstyles and choice of clothing are immensely variable. Challenges arise from various forms of facial adornment, such as piercings, jewellery, or glasses. Occlusions in the image can become quite serious, with only portions of the face visible, and in extreme cases, the face may not be visible at all.

When more of the body is depicted, the degrees of freedom available for the pose rise dramatically. More scope for accoutrements

becomes available: for example, an artist might be depicted holding a paintbrush, or an animal lover shown holding a parrot. In modern photography, an individual might be shown carrying out some characteristic action, or a photograph might seek to capture an intense moment, such as in a sporting competition.

Finally, greater freedom in the surrounding context can complicate stylisation. A portrait may gain some of its impact from the surroundings depicted, or some visual interest from an unusual arrangement, such as a face seen through glass and partly obscured by a reflection. Certainly, a portrait need not be restricted to a single individual; portraits of families are a contrary example, and modern selfie culture often produces photographs of small groups of friends.

As we ascend to the highest levels of difficulty, the complications can become extreme. Figure 2 shows examples of quite pronounced complications, including costumes, heavy occlusions, and multiple individuals. In the case of the drummer, the full effect of the portrait cannot be obtained by stylising only the human subject, as the context provides much of the impact. These examples are not even the most challenging images possible, but serve to illustrate the gulf between the first levels of the benchmark and the full range of photos people might seek to stylise.

Figure 2: Some difficult cases for dedicated portrait stylisation methods. All images came from pexels. Photo credits: Thibault Trillet, Alex Holt, Annie Spratt, Mantas Hesthaven.

3.1 Level 1

Since level 1 should provide images that are straightforward to stylise, many restrictions are imposed. Each image should contain only a frontal, approximately upright, and unoccluded view of a single face which has a forwards gaze direction. The images must contain essentially no background objects or clutter; we also exclude other body parts, such as the hands. The backgrounds are homogeneous, but natural – they were not manually masked out. Consequently, the images are dominated by the face, which should fill most of the image and be cropped approximately at the neck so as to include minimal clothing. To simplify the task of processing and rendering the face, we exclude both facial hair and long hair that partly covers the face. Also, the image should be free from "ornaments" such as a pipe, glasses, hat, or large pieces of jewellery. Harsh or complex lighting is avoided, and only soft lighting used. Given the above restrictions, as well as the other restrictions such as copyright and size, the majority of images used for this level of the benchmark will be posed portraits rather than impromptu snapshots. We have found that such portraits tend to have a limited range of expressions: either neutral or a moderate smile. We have therefore allowed all images at level 1 to have these expressions without requiring further control; variation in expression will appear in level 2.

Next, we specify the desirable variations. There should be a roughly equal distribution of gender. Given the variety of face shapes that occur, face shape should also be systematically controlled. We settled on the following set of folk descriptions of face shapes: {round, square, oval, heart, long}, although we note that these are not strictly defined, and this along with subtle differences between some shapes means that the attribution of face shape to images is only approximate. Degree of attractiveness, our next attribute, is also subjective. There is a tendency in the NPR literature to use aesthetically pleasing images with attractive and/or interesting faces. We have specified three levels of attractiveness: {less, average, more}. Finally, we aim to evenly cover three different ethnicities {white, asian, black} and two age groups: {young adult, middle-aged adult}. Again, when selecting images, we note that determining these attributions will be approximate and subjective. However, it is not critical that the image characteristics are precise, but rather that a reasonably uniform sampling is carried out.

3.2 Level 2

Level 2 contains many of the restrictions enforced in level 1: each image contains a frontal, approximately upright, unoccluded view of a single face that fills most of the image, is cropped to include minimal clothing, and does not include hands or other body parts. The background should be relatively plain, but since this requirement is not as strict as for level 1, some mostly unobtrusive background content is present. The requirement for unadorned faces is also relaxed, and so some hats and jewellery are allowed. Likewise, level 1's requirement for moderate lighting is maintained, but relaxed a little. Gaze direction is mostly forwards, but not exclusively. Ages are again restricted to adult, but are not considered as a control variable for this level.

Regarding desirable variations, like level 1 there should be a roughly equal distribution of gender. We would like to include different facial expressions, and it seemed reasonable to use Ekman's [Ekman 1972] universal expressions: happiness, sadness, anger, disgust, fear and surprise. However, when searching for images, we found it difficult to find sufficient examples that also satisfied the other level 2 conditions. In fact, with the exception of happiness, these expressions are not common in everyday conversations, and moreover, many other expressions that do naturally occur – thoughtfulness, agreement, confusion, and boredom, among others – are not included. Therefore, to expedite sourcing suitable images, we simplified the required range of facial expressions to the set of {neutral, positive, negative}. Furthermore, extreme versions of facial expressions should be avoided; not only could these pose a challenge for rendering, but often cause problems for the fitting of face models, often required for NPR portraits. The final factor to

control at level 2 is to include varieties of facial hair; we used the following categories: {none, moustache, beard, goatee, stubble}.

3.3 Design Matrix

Ideally, to populate the benchmark, we would like to systematically include images that provide examples of all the combinations of the portrait characteristics that we are controlling. However, a full factorial design would result in $2 \times 5 \times 3 \times 3 \times 2 = 180$ images, somewhat beyond what is feasible to present in a paper and difficult to manually evaluate. Thus, we will sample the combinations and limit the benchmark to a target of 20 images.

When Mould and Rosin [2016] created *NPRgeneral*, they manually selected a sample of 20 images to cover a range of desired characteristics. In this paper we take a more systematic approach, using the methodology of generating a "nearly orthogonal design matrix". This provides a representative subset of the values in the potentially high-dimensional input space of input conditions (variables) while maintaining approximate orthogonality of the input variables, which can be measured by computing the correlations between the input variables. We use the optFederov function from the R package AlgDesign [Wheeler 2014], which allows a number of runs (i.e. images) to be specified, as well as allowing for different numbers of values for each of the input variables. The resulting nearly orthogonal design matrices for levels 1 and 2 are shown in Tables 1 and 2.

gender	face shape	attractiveness	ethnicity	age
female	square	average	white	young
female	round	more	white	young
male	oval	more	white	young
male	square	average	black	young
male	long	average	black	young
male	oval	less	black	young
female	heart	more	black	young
female	long	average	asian	young
male	round	less	asian	young
female	heart	less	asian	young
male	heart	average	white	middle
female	square	less	white	middle
male	long	less	white	middle
male	round	average	black	middle
female	oval	average	black	middle
female	round	less	black	middle
female	long	more	black	middle
female	oval	average	asian	middle
male	square	more	asian	middle
male	heart	more	asian	middle

Table 1: Design matrix for level 1.

gender	expression	facial hair
male	negative	none
male	neutral	none
female	neutral	—
female	positive	—
male	negative	moustache
female	neutral	—
male	positive	moustache
female	positive	—
male	negative	beard
female	negative	—
male	neutral	beard
female	positive	—
female	negative	—
male	neutral	goatee
female	neutral	—
male	positive	goatee
female	negative	—
male	neutral	stubble
female	neutral	—
male	positive	stubble

Table 2: Design matrix for level 2.

3.4 Image Selection

Images matching the characteristics in the design matrices were sourced by the authors from online photography repositories, principally Flickr. As in Mould and Rosin's [Mould and Rosin 2016] NPR benchmark data set, we considered only images whose license permits distribution of modified versions. Further, we enforced the constraint that the images should be large enough so that the image height could be standardised at 1024 pixels, even after cropping out the face.

It is preferable to obtain the images from a wide variety of photographers, rather from a single image collection as is common in computer vision face databases. We wish to ensure that a variety of cameras, lighting conditions, backgrounds and poses are included. This will present a greater challenge to stylisation algorithms, as well as providing a more representative cross-section of images.

We note that the design matrix specifications required the authors to estimate characteristics (face shape, attractiveness, ethnicity, age, expression) which are not always clear from the image, and in some cases also involved subjective judgements. However, as noted before it is not critical that the image characteristics are precise, but rather that a reasonably uniform sampling is carried out.

Following this process we selected 20 images according to the design matrices. This was more difficult than expected, since the majority of Flickr photographs are taken under uncontrolled conditions, and hence have complicated backgrounds, harsh lighting, non-frontal view, occlusion or other factors that forced us to reject them. The level 1 and 2 images can be found in Figure 3 and Figure 4 respectively.

4 NPR ALGORITHMS

We applied six stylisation methods to the benchmark set. Discussion of the outcomes is found in the following section. Here, we give an overview of each of the methods and describe the parameter settings employed in generating the results.

Rosin and Lai's algorithm [Rosin and Lai 2015] first stylises the image with abstracted regions of flat colours plus black and white

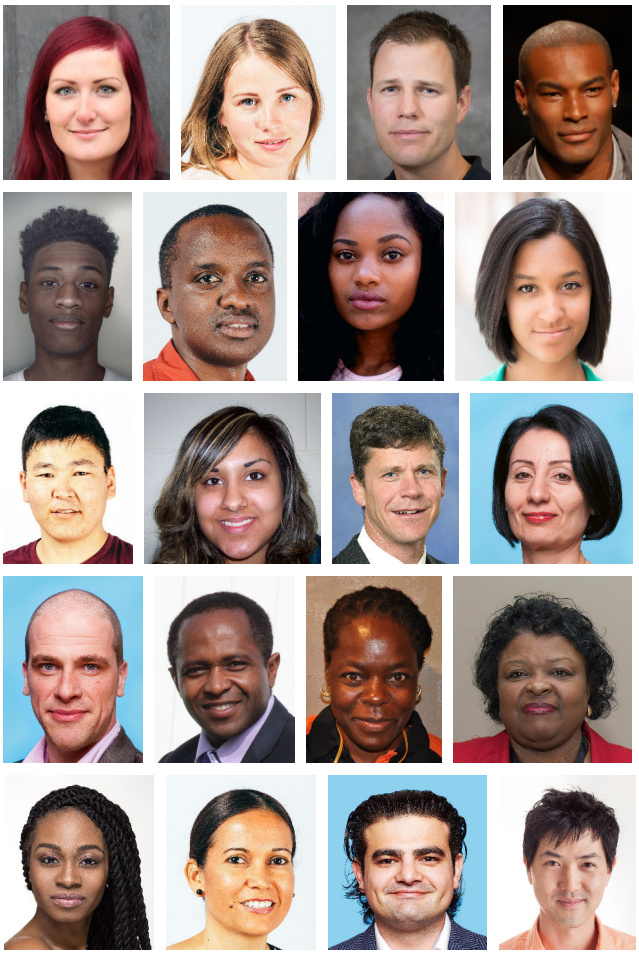

Figure 3: Source images comprising level 1 of the portrait benchmark. All images came from Flickr unless otherwise specified. Photo credits: Pirátská strana, IFES - International Fellowship of Evangelical Students, Mecklenburg County, Jesse Gross (Wikimedia), iKobe, IFES - International Fellowship of Evangelical Students, Pexels (pixabay), Ethan M Sigmon, IFES - International Fellowship of Evangelical Students, projectofheart, Oregon Department of Forestry, Partij van de Arbeid, Partij van de Arbeid, chidi (pixabay), SANGONeT ICT for NGOs Conference, Mecklenburg County, jaymarable, IFES - International Fellowship of Evangelical Students, Partij van de Arbeid, Matthew Roth.

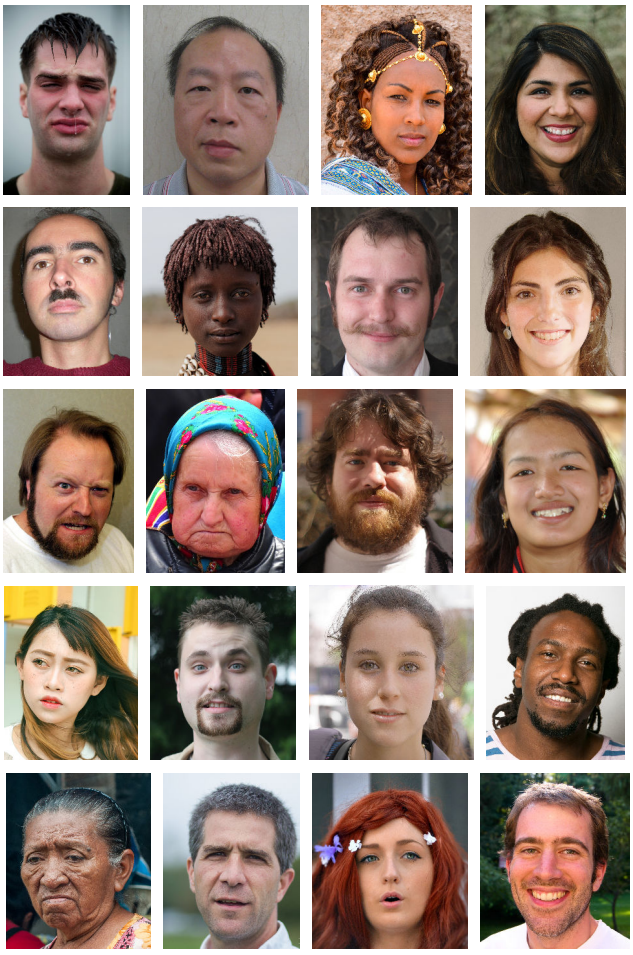

Figure 4: Source images comprising level 2 of the portrait benchmark. All images came from Flickr unless otherwise specified. Photo credits: Sgt. Matthew Callahan (Wikimedia), BBC World Service, Rod Waddington, Adam McGuffie, Pablo El Diablo, www.j-pics.info, susan, Nando.uy, Christopher Blizzard, Adam Jones, Christopher Thompson, shankar s., ptksgc (pixabay), Sparky, Nando.uy, Martin Sharman, wiki The Photographer, Greg Peverill-Conti, Hamish Irvine, Greg Peverill-Conti.

lines [Lai and Rosin 2014], then fits a partial face model to the input image and attempts to detect the skin region. Shading and line rendering is stylised in the skin region, and in addition, the face model helps inform portrait-specific enhancements: reducing line clutter; improving eye detail; colouring the lips and teeth; and inserting synthesised highlights.

Wang et al. [Wang et al. 2013] proposed an example-based rendering technique that learns a non-parametric model of style by observing the geometry and tone of brush strokes in an exemplar photo-painting pair. The novelty of the approach is in modulating stroke attributes directly rather than pixel patches (as with Image Analogies [Hertzmann et al. 2001]) to render by example, and the technique is specialised to portraiture by learning the stroke models within independent semantic regions of the face. The algorithm uses a Markov Random Field (MRF) model to ensure spatial coherence of stroke style during learning and rendering.

Berger et al. [Berger et al. 2013] mimic the style of specific artists' line-drawings in a data-driven manner. Sample drawings of artists are collected and their statistics are analysed. Then, given a new portrait photograph and an artist style, the algorithm first creates a contour image by using a variant of the XDoG method [Winnemöller et al. 2012]. Using the detected facial features, the face geometry is modified to follow the specific artist's geometric style.

Lastly, the face contours are drawn using strokes from the artist's stroke database following the artist's drawing statistics.

Li and Wand's method [Li and Wand 2016] treats styles as textures, and forces the synthesised image and the reference style image to have the same Markovian texture statistics. Non-parametric sampling is first used to capture patches from the style image; patch matching and blending are then used to transfer the style to the synthesised image. For portrait stylisation, they include an additional content constraint that minimises the L_2 distance between the CNN encoding of the portrait photo and the synthesised image. Their method reduces implausible feature mixtures that are common to previous CNN based approaches, permitting synthesizing photographic content with increased visual plausibility. However, the method can be too rigid for some painterly styles, generates artefacts due to mismatch between the content and the style images, and requires hundreds of rounds of iterations to achieve good results.

Winnemoeller et al.'s XDoG filter [Winnemöller et al. 2012] can be conceptualised as the weighted sum of a blurred source image and a scaled difference-of-Gaussians (DoG) response of the same image, effectively applying unsharp masking to the DoG response. Combined with subsequent soft thresholding, this computationally simple filter allows a wide range of stylistic and artistic effects, including cartoon shading, black-and-white thresholding, and charcoal shading. Faces are handled well, despite not being treated differently from other image content. If required, local modification of filter parameters, according to facial features, would be trivial to implement.

Li and Mould's stippling method [Li and Mould 2011] is an adaptation of contrast-aware halftoning [Li and Mould 2010], an error diffusion method where pixels are thresholded in a priority order and the resulting error distributed so as to preserve contrast: negative error goes preferentially to darker pixels, and positive error to lighter pixels. For stippling, thresholding down corresponds to placing a stipple, and thresholding up means placing nothing. The method strives to maintain the original intensity level while preserving local details.

The full set of level 1 benchmark images processed by each method are shown in Figures 5 through 10. Several of the methods are capable of producing more than one output style; we show some alternative styles in Figure 11. For these figures, and indeed all images shown in the paper, we urge the reader to view the images at full resolution in electronic form: some small yet important details are not fully apparent on the page or at low resolution.

Parameter settings were as follows:

- Painterly portraits [Wang et al. 2013] used fixed parameters, with the same values as were used in the original paper.
- Rosin and Lai used fixed parameter settings over the benchmark, as detailed in the original paper [Rosin and Lai 2015], with an additional check: dark and light faces were differentiated by testing whether the mean intensity of a central region in the face lies below the mid intensity range value (128). In HSI space, intensity values for three classes of pixels are normally quantised to {0,200,255}. For dark faces, the mean intensity value for each class was used instead.

Figure 5: Portrait benchmark images stylised using the example-based painterly method of Wang et al.

- Portrait sketching [Berger et al. 2013] did not do any parameter tuning for these results. However, note that the results are artist-dependent, since statistics of the specific artist are used to create the results.
- The style transfer method [Li and Wand 2016] used the following parameter settings: patch size of 3×3 on both layer *relu3_1* and *relu4_1* for the style constraint. The selection of the layers is based on empirical study of the matching and blending performance of different layers. In general, larger patches preserve more features from the style image at the cost of being increasingly rigid; smaller patches can adapt to the content image more easily but have the risk of losing characteristic meso-structures.
- Stippling [Li and Mould 2011] used fixed parameters for all images: exaggeration coefficients $G_+ = 5$ and $G_- = 5$, base stipple size $k = 0.1$, and mask size $D = 15$. The method executed error diffusion over a raster of height 512.
- XDoG [Winnemöller et al. 2012] used mostly fixed parameters. However, since the XDoG contains a luminance

Figure 6: Portrait benchmark images stylised using the line and region method of Rosin and Lai.

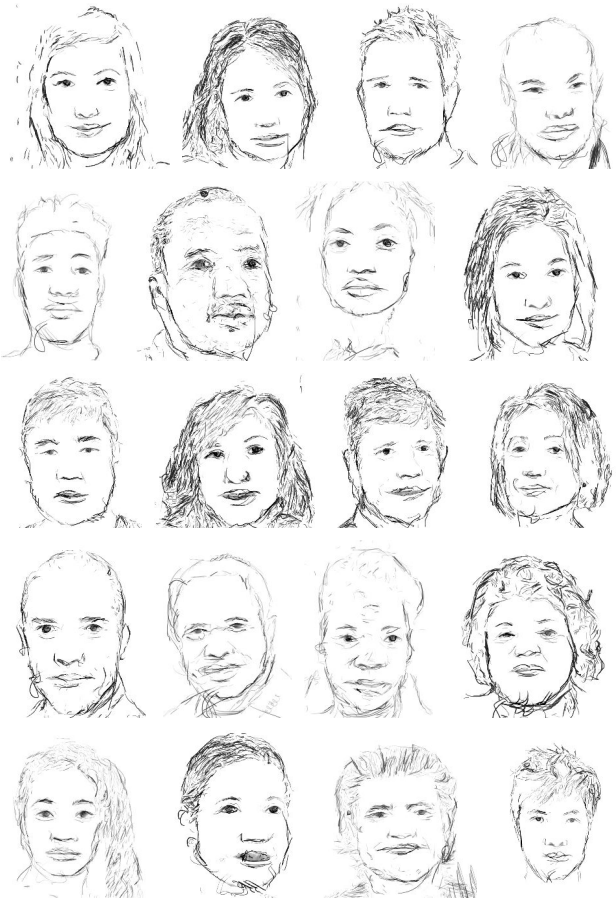

Figure 7: Portrait benchmark images stylised using the portrait sketching method of Berger et al.

thresholding operation, we used two sets of tone-mapping parameters: the values for $\{p, \epsilon_p, \phi_p\}$ were set to $\{17, 69.5, 0.03\}$ for images with predominantly dark tones, for images with lighter tones, $\{46.7, 79.5, 0.017\}$. The DoG parameters were $\sigma_e = 1.395$, $\text{dog}_k = 1.6$, $\sigma_m = 4$, and 4 iterations of ETF smoothing were used.

5 DISCUSSION

The portrait algorithms are generally successful at treating the benchmark images: they are all able to consistently stylise the images, and did not suffer any major breakdowns over the benchmark's fixed set of images. Though the subjects may not always be identifiable as individuals, facial elements are generally clear.

General NPR algorithms, here exemplified by XDoG and stippling, also produce recognisable images. In the case of the relatively straightforward inputs at level 1, there seems to be little quality difference between general methods and portrait-specific methods. Put another way, the benefits of the face-specific elements of the dedicated portrait methods do not seem very strong.

However, face-specific algorithms do have advantages. Rather than simply re-rendering the input image using low-level image processing operations, the face-specific methods have an underlying model of the face, allowing them to make more elaboration abstractions (Berger et al.) that would be infeasible in the absence of such a model, or improving robustness (Rosin and Lai) by fixing problems that show up in the low-level processing.

Each method benefits from the face-specific information in a different way. Rosin and Lai use the fitted face to add highlights and shading; the synthetic nose compensates for the faint features in image 2. Berger et al. perform shape abstraction by modifying the geometry of the face according to the artist's style and placing strokes based on a distribution obtained from hand-drawn sketches of faces. Wang et al. ensure feature preservation and are able to improve the contrast between the faces and the backgrounds. Finally, the CNNMRF method is able to propagate styles while largely preserving face structure owing to their MRF prior that encourages a layout consistent with the provided face image. However, due to the VGG model's prior training for general object detection, and the

Figure 8: Portrait benchmark images stylised using the CNN MRF method of Li and Wand.

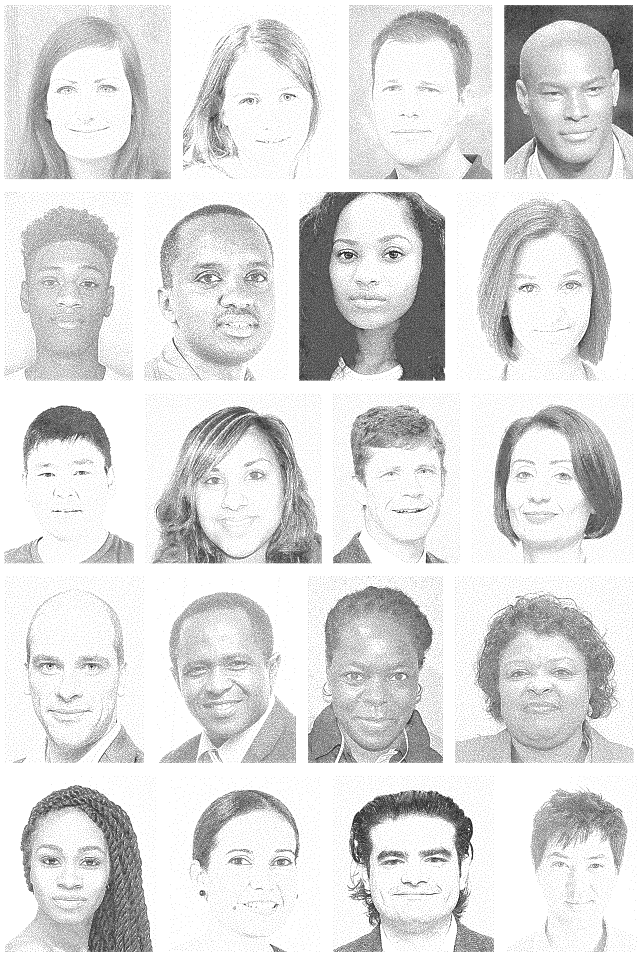

Figure 9: Portrait benchmark images stylised using the structure-preserving stippling method of Li and Mould.

discriminative nature of eyes, there is a tendency for the CNNMRF method to include spurious eyes in the portraits.

Problems do appear in the stylised images. Rarely are significant details omitted, but image elements can be distorted. The painterly rendering system has presented the nose in image 2 in a confused way. The sketchy portraits are intentionally slightly distorted, but sometimes the jaw and mouth regions are excessively changed, as in images 2, 14, and 16. Teeth, in particular, are sometimes a problem: both the sketched portraits system and the CNNMRF system fill in open mouths in an unpleasant-looking way.

The most widespread problem is the addition of spurious material, appearing in all methods to greater or lesser degree. The painterly strokes sometimes discolour the face, as in images 6 and 13. Rosin and Lai's method adds significant discolourations to a few images, especially images 1 and 7, and stray lines sometimes clutter the results, such as in images 12 and 19. The sketched portraits have stray lines as well, especially around the chin region; image 6 has a long erroneous stroke along the man's cheek. The CNNMRF method has a tendency to add eyes in unlikely locations; in the Picasso style, it sometimes adds unnecessary bold lines across the face.

This litany of problems might seem to support the naive view that portrait-specific methods do not produce obviously superior stylisations to more general low-level algorithms. However, the general methods have problems too. In the XDoG results, several face outlines are broken or weakly shown; in a few cases, such as images 9 and 18, the gaps are very large. XDoG is sensitive to the choice of threshold, and despite generally controlling for image brightness, the face areas can be patchy. Images 6 and 16 show the problem: we would like to convey the subjects' dark skin, but neither choosing a high threshold (and letting the image become mostly white) nor choosing a lower one (and letting the skin colour vary) are entirely satisfactory. In the stippling results, weak boundaries are not always clear, and an overall policy of greylevel preservation leads to distracting stippling coverage of the background regions.

Additional examples of general stylisation methods can be found in Figure 12. These methods usually worked well on the benchmark

Figure 10: Portrait benchmark images stylised using the XDoG method of Winnemöller et al.

images, but were occasionally less successful. For each method, we show both a typical result (above) and a flawed output (below).

With its portrait-specific aspects disabled, Rosin and Lai's method has several mild shortcomings. Faces tend to be cluttered with distracting small lines, and the eyes are often fringed with white. In some cases (right example) the clutter from lines and colour quantisation can be severe. Gastal and Oliveira's abstraction method is successful much of the time, but important details might lack contrast and hence fade, while spurious details such as lighting changes might be enhanced; the eyes and teeth in the right-hand result provide examples. Similarly, the color shifting of Sisley the abstract painter works well, calling to mind images from Andy Warhol; however, low-contrast details are not preserved and even high-contrast elements such as the eyes may not be presented properly in the output, depending on the vagaries of the random strokes. The variant of Hertzmann's method rarely fails badly, but small or low-contrast details may not be preserved depending on the settings. The same comments apply to the circular scribble

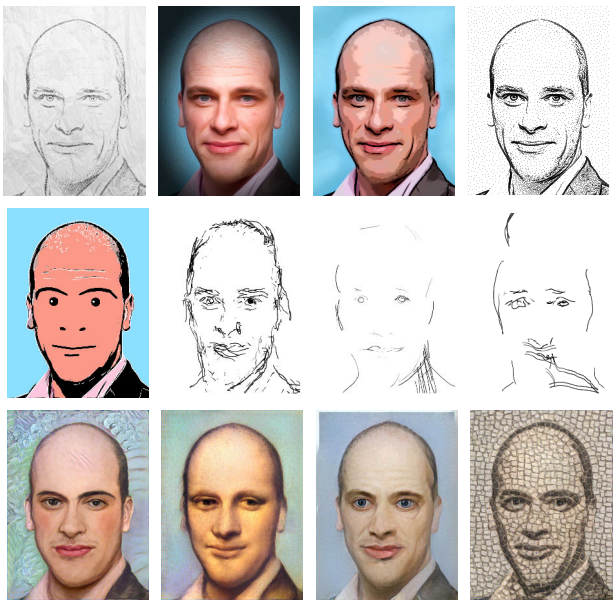

Figure 11: Portrait benchmark image rendered in alternate styles. Top: XDoG hatching; XDoG oil paint; XDoG toon; stippling guided by ETF lines. Middle: Julian Opie variant of lines and blocks style; sketchy portrait with alternate artist; sketchy portrait with high abstraction; high abstraction and alternate artist. Bottom: CNNMRF variants: Frida Kahlo, *Mona Lisa*, James Hague, and mosaic styles.

method: large regions of high contrast are fine, but small details and low-contrast boundaries are omitted.

Figure 13 shows some results from existing portrait stylisation results applied to level 2 of the benchmark image set. The results are mixed. Since level 2 is only a small increment in difficulty, output quality is not enormously worse. Nonetheless, the new complications do challenge the methods. Rosin and Lai's results suffer from increased clutter, and facial expressions can prevent the synthetic lips from integrating well into the image. Wang et al.'s method fares well, nicely portraying the beard in the third image, but the face shape and expression in the second image and the facial hair in the fourth image are not entirely clear. Li and Wand's transferred Picasso style is still effective, but the incidence of small defects is higher and the method does not seem to cope well with facial hair. In general, facial expressions and visible teeth will pose problems for methods that make more confining assumptions about mouth pose.

The general methods will remain effective on more difficult face images, while many face-specific methods will be overcome by the complications as the input becomes less constrained. Thus, in part this paper is a call to arms asking those interested in face stylisation to take up the challenge and try to do more.

We are hopeful that the benchmark image set will help with this endeavour. The current two levels provide a range of faces to exercise future methods, with the second level demonstrating more complications than have traditionally been attempted by many

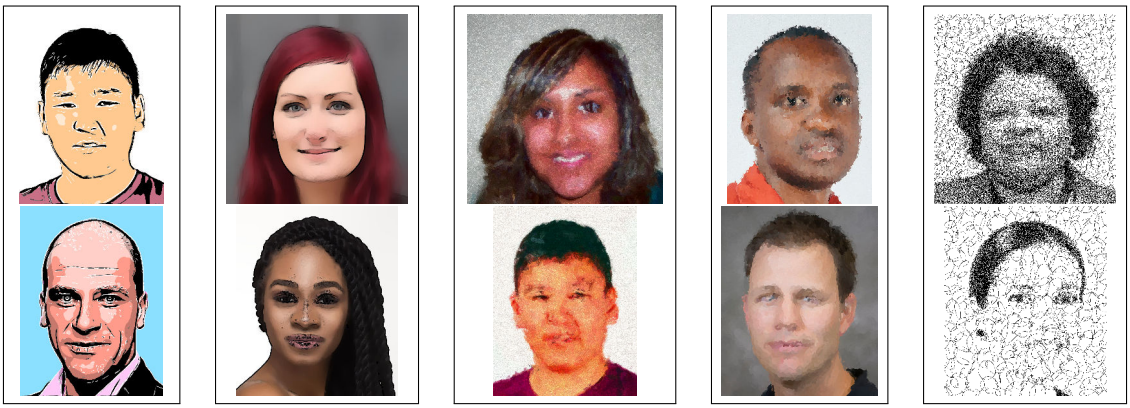

Figure 12: Level 1 benchmark images processed by general NPR stylisation methods. First pair: Rosin and Lai's algorithm [Rosin and Lai 2015] without the face model. Second pair: Gastal and Oliveira's method [Gastal and Oliveira 2011]. Third pair: Sisley the abstract painter [Zhao and Zhu 2010]. Fourth pair: Variant of Hertzmann's layered painterly stylisation [Hertzmann 1998]. Bottom pair: Circular scribble art [Chiu et al. 2015].

automated portrait stylisation techniques. Even the first set, with its mild complications such as teeth, its varied contrast levels, and its range of face shapes, has been informative in understanding the capabilities of existing portrait methods. Further, the results we show here should be helpful to future researchers in evaluating their methods and comparing against past work.

6 CONCLUSION AND FUTURE WORK

In this paper, we have presented two levels of a specialised benchmark to aid in evaluating portrait stylisation algorithms. The first level corresponds to the highly constrained portrait images usually used by existing portrait-specific stylisation techniques, with closely cropped faces in strictly frontal view, clean backgrounds, no facial hair or ornamentation, and neutral or mildly positive facial expressions. The second level relaxes the constraints on pose, lighting, and background slightly, while adding the complications of facial hair and more varied expressions. Both sets contain an even distribution of genders and a range of ethnicities. The intent is that researchers can apply their algorithms to these image sets, facilitating comparisons and ensuring neutrality in image selection. We applied six existing stylisation algorithms to the level 1 benchmark images: three existing portrait-specific stylisation algorithms, two general-purpose stylisation algorithms, and one general learning based stylisation algorithm. The general methods coped well, but the domain knowledge from the face-specific methods enabled them to improve the quality and robustness of their stylisation. Conversely, when we advanced to level 2, there was no change in the effectiveness of the general methods, but the additional complications posed challenges to the face-specific methods: in particular, facial hair sometimes confounded the algorithms.

The current benchmark presented in this paper should be considered as a basic "version 0.1". Future work will look at performing user studies to confirm the separation in difficulty between levels 1 and 2, and between level 2 and the proposed levels 3 and 4, which we have proposed but not constructed. The third and fourth levels will further relax the constraints on expression, background, pose, and lighting, while extending the range of subjects to include children and the elderly. Other complications reserved for yet higher levels include more elaborate poses, full-body photographs, photos of multiple people, partial occlusions, and unconstrained backgrounds.

Further future work can involve both formally extending the benchmark to even higher levels, as well as further employing the benchmark's existing two levels. We have used four portrait-specific methods in this paper, but many more exist, and it would be of interest to document the outcomes from applying additional existing methods to the benchmark set.

Of course, future work need not be limited to work on the benchmark per se. On the contrary, the most interesting directions of future work are about portrait stylisation itself. The benchmark is intended as a tool to help with evaluation of future stylisation methods, making comparisons easier and more systematic. Finally, we hope that this paper spurs further and more ambitious work, by arguing that existing portraiture methods in NPR are highly constrained both compared to the range of photographs that people take and compared with the historical practice of portraiture.

ACKNOWLEDGEMENTS

Thanks to the reviewers for helpful comments. Particular thanks to the various photographers, named and anonymous, who provided images. This work was supported in part by NSERC.

REFERENCES

Brian Becker and Enrique Ortiz. 2013. Evaluating open-universe face identification on the web. In *Proc. Analysis and Modeling of Faces and Gestures Workshop*. 904–911.

Itamar Berger, Ariel Shamir, Moshe Mahler, Elizabeth Carter, and Jessica Hodgins. 2013. Style and abstraction in portrait sketching. *ACM Transactions on Graphics (TOG)* 32, 4 (2013), 55.

Susan E Brennan. 2007. Caricature generator: The dynamic exaggeration of faces by computer. *Leonardo* 40, 4 (2007), 392–400.

Chun-Chia Chiu, Yi-Hsiang Lo, Ruen-Rone Lee, and Hung-Kuo Chu. 2015. Tone- and Feature-Aware Circular Scribble Art. *Computer Graphics Forum* 34, 7 (2015), 225–234.

Simon Colton, Michel F. Valstar, and Maja Pantic. 2008. Emotionally Aware Automated Portrait Painting. In *Int. Conf. on Digital Interactive Media in Entertainment and Arts*. 304–311.

Figure 13: Selected level 2 benchmark images processed by portrait stylisation methods. Top row: Rosin and Lai's method. Second row: Wang et al.'s method. Third row: Berger et al.'s method. Fourth row: Li and Wand's method. Fifth row: Li and Mould's method. Bottom row: Winnemöller et al.'s method.

Paul Ekman. 1972. Universal and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation*. 207–283.

Wolfgang Förstner. 1996. 10 Pros and Cons Against Performance Characterization of Vision Algorithms. In *Workshop on Performance Characteristics of Vision Algorithms*.

Brendan Galea, Ehsan Kia, Nicholas Aird, and Paul G Kry. 2016. Stippling with aerial robots. In *Proc. of the Joint Symposium on Computational Aesthetics and Sketch Based Interfaces and Modeling and Non-Photorealistic Animation and Rendering*. 125–134.

Eduardo S. L. Gastal and Manuel M. Oliveira. 2011. Domain Transform for Edge-aware Image and Video Processing. *ACM Trans. Graph.* 30, 4, Article 69 (2011), 12 pages.

Paul Haeberli. 1990. Paint by numbers: Abstract image representations. In *ACM SIGGRAPH*, Vol. 24. 207–214.

Peter Hall and Ann-Sophie Lehmann. 2013. Don't Measure – Appreciate! NPR Seen Through the Prism of Art History. In *Image and Video-Based Artistic Stylisation*, Paul L. Rosin and John P. Collomosse (Eds.). Springer, 333–351.

Robert M. Haralick. 1994. Performance Characterization in Computer Vision. *CVGIP: Image Understanding* 60, 2 (1994), 245–249.

Aaron Hertzmann. 1998. Painterly rendering with curved brush strokes of multiple sizes. In *ACM SIGGRAPH*. ACM, 453–460.

Aaron Hertzmann. 2010. Non-Photorealistic Rendering and the Science of Art. In *Proceedings of the 8th International Symposium on Non-Photorealistic Animation and Rendering*. 147–157.

Aaron Hertzmann, Charles E. Jacobs, Nuria Oliver, Brian Curless, and David H. Salesin. 2001. Image Analogies. In *ACM SIGGRAPH*. 327–340.

Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. 2007. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments.* Technical Report 07-49. University of Massachusetts, Amherst.

Tobias Isenberg. 2013. Evaluating and Validating Non-photorealistic and Illustrative Rendering. In *Image and Video-Based Artistic Stylisation*, Paul L. Rosin and John P. Collomosse (Eds.). Springer, 311–331.

Yu-Kun Lai and Paul L. Rosin. 2014. Efficient Circular Thresholding. *IEEE Trans. Image Processing* 23, 3 (2014), 992–1001.

Helmut Leder, Benno Belke, Andries Oeberst, and Dorothee Augustin. 2004. A model of aesthetic appreciation and aesthetic judgments. *British Journal of Psychology* 95, 4 (2004), 489–508.

Chuan Li and Michael Wand. 2016. Combining Markov Random Fields and convolutional neural networks for image synthesis. In *Proc. Computer Vision and Pattern Recognition*. 2479–2486.

Hua Li and David Mould. 2010. Contrast-aware Halftoning. *Computer Graphics Forum* 29, 2 (2010), 273–280.

Hua Li and David Mould. 2011. Structure-preserving stippling by priority-based error diffusion. In *Proceedings of Graphics Interface 2011*. Canadian Human-Computer Communications Society, 127–134.

Michael J Lyons, Julien Budynek, and Shigeru Akamatsu. 1999. Automatic classification of single facial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21, 12 (1999), 1357–1362.

David Mould. 2014. Authorial Subjective Evaluation of Non-photorealistic Images. In *Proceedings of the Workshop on Non-Photorealistic Animation and Rendering (NPAR)*. 49–56.

David Mould and Paul L. Rosin. 2016. A benchmark image set for evaluating stylization. In *Proceedings of the Joint Symposium on Computational Aesthetics and Sketch Based Interfaces and Modeling and Non-Photorealistic Animation and Rendering*. 11–20.

Sven C. Olsen and Bruce Gooch. 2011. Image simplification and vectorization. In *Proc. ACM Symposium on Non-photorealistic Animation and Rendering*. 65–74.

Stephen E Palmer, Karen B Schloss, and Jonathan Sammartino. 2013. Visual aesthetics and human preference. *Annual Review of Psychology* 64 (2013), 77–107.

P.J. Phillips, Hyeonjoon Moon, S.A. Rizvi, and P.J. Rauss. 2000. The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 10 (2000), 1090–1104.

Paul L Rosin and Yu-Kun Lai. 2015. Non-photorealistic rendering of portraits. In *Proceedings of the Workshop on Computational Aesthetics*. Eurographics Association, 159–170.

Ferdinando S Samaria and Andy C Harter. 1994. Parameterisation of a stochastic model for human face identification. In *Proc. Workshop on Applications of Computer Vision*. 138–142.

Ahmed Selim, Mohamed Elgharib, and Linda Doyle. 2016. Painting Style Transfer for Head Portraits Using Convolutional Neural Networks. *ACM Trans. Graph.* 35, 4 (2016), 129:1–129:18.

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *Proc. Conference on Computer Vision and Pattern Recognition*. 1701–1708.

Neil A Thacker, Adrian F Clark, John L Barron, J Ross Beveridge, Patrick Courtney, William R Crum, Visvanathan Ramesh, and Christine Clark. 2008. Performance characterization in computer vision: A guide to best practices. *Computer vision and image understanding* 109, 3 (2008), 305–334.

Tinghuai Wang, John P Collomosse, Andrew Hunter, and Darryl Greig. 2013. Learnable Stroke Models for Example-based Portrait Painting. In *BMVC*.

Nicholas Westlake, Hongping Cai, and Peter Hall. 2016. Detecting People in Artwork with CNNs. In *Workshop on Computer Vision for Art Analysis*. 825–841.

Bob Wheeler. 2014. AlgDesign: Algorithmic Experimental Design. R Package Version 1.1-7. (2014).

Holger Winnemöller, Jan Eric Kyprianidis, and Sven C. Olsen. 2012. XDoG: An eXtended difference-of-Gaussians compendium including advanced image stylization. *Computers & Graphics* 36, 6 (2012), 740–753.

Mingtian Zhao and Song-Chun Zhu. 2010. Sisley the Abstract Painter. In *ACM Symp. NPAR*. 99–107.

Mingtian Zhao and Song-Chun Zhu. 2013. Abstract painting with interactive control of perceptual entropy. *ACM Transactions on Applied Perception (TAP)* 10, 1 (2013), 5.