Complexity

Complex Systems in Aesthetics and Arts

Lead Guest Editor: Juan Romero
Guest Editors: Colin Johnson and Jon McCormack

Complex Systems in Aesthetics and Arts

Complexity

Complex Systems in Aesthetics and Arts

Lead Guest Editor: Juan Romero
Guest Editors: Colin Johnson and Jon McCormack

Editorial Board

Contents

Hindawi
Complexity
Volume 2019, Article ID 9836102, 2 pages
https://doi.org/10.1155/2019/9836102

Editorial

Complex Systems in Aesthetics and Arts

Juan Romero[1], **Colin Johnson**[2], **and Jon McCormack**[3]

[1] *University of A Coruña, A Coruña, Spain*
[2] *University of Kent, Canterbury, UK*
[3] *Monash University, Melbourne, Australia*

Correspondence should be addressed to Juan Romero; jj@udc.es

Received 16 April 2019; Accepted 16 April 2019; Published 30 April 2019

The arts are one of the most complex of human endeavours, and so it is fitting that a special issue on *Complex Systems in Aesthetics and Arts* is being published. As the editors of this special issue, we would like to thank the reviewers of the submitted papers for their hard work in making this issue possible, as well as the authors who submitted their work and were very responsive to the comments of the reviewers and editors.

The word *complexity* has a specific meaning in the context of "complex systems" research, as the study of systems made of many components—not in themselves necessarily complex—that through loosely coupled, local interactions generate complex, emergent behaviours. Such systems have the potential to act as the basis for the production of artworks, whether entirely computer generated or as a result of a cocreative system between humans and computers. Such art might make its impact through the intrinsic interest of the complex behaviour in the system, by representing, exploring, or connoting some worldly aspect of complexity, or by using complex systems as a way of exploring a space of possible works. Furthermore, complex systems research has the potential to simulate emergent processes in the artworld, such as the interaction between artists, audiences, and critics, or the development of aesthetic ideas or artistic fashions over time.

The context for the special issue is explored in the first paper, *Understanding Aesthetics and Fitness Measures in Evolutionary Art Systems*, authored as an overview paper on the topic by the editors and I. Santos. This takes a particular algorithm that is grounded in complexity science ideas—evolutionary search—and explores links between the construction of fitness measures in these systems, and measures and concepts of aesthetic value from the philosophy and psychology of art. A common feature of complex systems is that individual agents make evaluations as a driver for behaviour, and so the links formed in this paper between the human behaviour of making aesthetic judgements and similar processes in computer systems have the potential to inform work in many applications of complex systems to the arts.

This theme of aesthetic measures is continued in the paper by A. Carballal et al., *Avoiding the Inherent Limitations in Datasets Used for Measuring Aesthetics When Using a Machine Learning Approach.* In this paper, the authors explore how well a machine learning approach can replicate the aesthetic and quality judgements of a number of humans across a large set of photographs, exploring whether the machine learning algorithms can learn to replicate and generalise from human judgements and then apply these accurately to new examples. The paper also addresses whether the learned models replicate the phenomenon found in the human results whereby aesthetic value and technical quality are correlated. They conclude that the correlation is present also in the learned models, though less strongly than with humans, and that the machine learning models were typically better at assessing (more objective) technical quality than (more subjective) aesthetic value.

The remaining articles explore a variety of other topics concerned with aesthetic aspects of images and graphics. *Evolutionary Computation for Modelling Social Traits in Realistic Looking Synthetic Faces* by F. Fuentes-Hurtado and colleagues explores the use of evolutionary computation to select sets of facial features that convey a particular social emotion and then uses an automated image editing approach, Poisson Image Editing, to create a realistic composite image that combines the chosen features. By contrast with the realistic images in that paper, *Image Evolution Using 2D Power Spectra* by M. Gircys and B. J. Ross uses evolutionary algorithms to produce abstract artworks based on realistic photographs and paintings. The system is based on a spectral analysis of the original image, which is used to construct a fitness function that then drives the evolutionary process to generate novel images based on the same spectral profile. The system produces images that still connote features of the source image but are more abstract.

Finally, the paper *Evolving Stencils for Typefaces: Combining Machine Learning, User's Preferences and Novelty*, by T. Martins et al., explores a system with two components. The first of these is an evolutionary system for exploring the complex search space of typefaces. The second component is a human-computer cocreative system to develop the fitness function that is used by the evolutionary algorithm. The paper demonstrates an exemplary piece of work in combining human and computer expertise in a complex aesthetic domain.

We believe that this selection of articles offers an interesting and timely insight into the interactions between aesthetics, machine learning, and computational creativity. We hope that you enjoy and learn from reading the papers in this issue.

Conflicts of Interest

The editors have no conflict of interest regarding the publication of this special issue.

Juan Romero
Colin Johnson
Jon McCormack

Hindawi
Complexity
Volume 2019, Article ID 3509263, 16 pages
https://doi.org/10.1155/2019/3509263

Research Article

Evolving Stencils for Typefaces: Combining Machine Learning, User's Preferences and Novelty

Tiago Martins, João Correia, Ernesto Costa, and Penousal Machado

CISUC, Department of Informatics Engineering, University of Coimbra, 3030 Coimbra, Portugal

Correspondence should be addressed to Tiago Martins; tiagofm@dei.uc.pt

Received 19 August 2018; Revised 28 January 2019; Accepted 4 March 2019; Published 26 March 2019

Guest Editor: Jon McCormack

Typefaces have become an essential resource used by graphic designs to communicate. Some designers opt to create their own typefaces or custom lettering that better suits each design project. This increases the demand for novelty in type design, and consequently the need for good technological means to explore new thinking and approaches in the design of typefaces. In this work, we continue our research on the automatic evolution of glyphs (letterforms or designs of characters). We present an evolutionary framework for the automatic generation of type stencils based on fitness functions designed by the user. The proposed framework comprises two modules: the evolutionary system, and the fitness function design interface. The first module, the evolutionary system, operates a Genetic Algorithm, with a novelty search mechanism, and the fitness assignment scheme. The second module, the fitness function design interface, enables the users to create fitness functions through a responsive graphical interface, by indicating the desired values and weights of a set of behavioural features, based on machine learning approaches, and morphological features. The experimental results reveal the wide variety of type stencils and glyphs that can be evolved with the presented framework and show how the design of fitness functions influences the outcomes, which are able to convey the preferences expressed by the user. The creative possibilities created with the outcomes of the presented framework are explored by using one evolved stencil in a design project. This research demonstrates how Evolutionary Computation and Machine Learning may address challenges in type design and expand the tools for the creation of typefaces.

1. Introduction

Typefaces are an essential resource employed by graphic designers [1], who are always willing to experiment with type and to explore new thinking, tools, and techniques. However, the creation of a typeface is a laborious process, involving the design of several glyphs for different characters. In the domain of type design, a glyph consists in a particular design of a character, e.g., a letter, figure, or punctuation mark. This, along with the increasing demand for new type design work, increases the need for good technological means to assist the designer in the creation of a typeface.

Although conventional computational design tools are effective for precise design tasks during the later phases of the design process, they offer insufficient support to design exploration during the earliest, essentially conceptual, stages of the design process. We also consider that most of the prominent software design tools tend to bias and limit the designers, who become accustomed to work and think in terms of the primitives that these tools provide, the workflow they induce, and the boundaries, implicit or explicit, that they establish. As a result, the outcome of the design project tends to be, at least partially, shaped by the tools, leading to visual tendencies. Therefore, we argue that it is as important to master and exploit the tools at hand, as it is to challenge those tools, by modifying them or inventing new ones that suit unique ideas and design projects.

In this work, we explore an evolutionary approach for the computational generation of glyphs. This approach is intended to provide the designer with a wide range of alternative designs as stimuli for inspiration, working in a mind-opening way and promoting new ideas. We do not expect our approach to competing with more traditional type design approaches, or to replace the designer. Our goal is to develop a tool that aids the designer.

Although some evolutionary approaches for type design exist [2–8] most of them rely on user evaluation, *i.e.* make use of Interactive Evolutionary Computation (IEC). Although asking the users to evaluate the designs being evolved enables them to directly influence the course of the evolution, this approach puts a considerable burden on them. This leads to user fatigue and, consequently, to the inefficient exploration of the search space. In addition, some of the identified evolutionary approaches require pre-existing typefaces or skeletons, the drawing of initial seed glyphs, or the identification of letter parts.

In the work *Evotype,* we have been combining Evolutionary Computation (EC) and Machine Learning (ML) to evolve glyphs in an autonomously way, with automatic fitness assignment. We started with a Genetic Algorithm (GA) that evolved different populations of candidate glyphs, one per target character, with and without migration of glyphs between populations [9, 10]. Although these early approaches were already able to evolve glyphs with expressiveness and legibility, the glyphs often lacked coherence. In other words, the evolved glyphs had no common visual structure and for this reason, they did not seem to be part of a single typeface. We addressed this limitation by evolving one population of individuals, each being able to express all the glyphs [11]. Each individual consists of a stencil composed of lines that can be used to construct glyphs. This approach provided more coherence to the final glyphs, since their structure share elements of the stencil. The fitness assignment was autonomous, and it was able to guide evolution towards stencils that produce simple, legible and coherent glyphs.

In this paper, we expand our approach to type stencil design by:

(i) Developing an ML approach to evaluate the glyphs produced by evolved stencils, combining a Convolutional Neural Network (CNN) with Self-Organising Maps (SOMs) to evaluate their recognisability as the target character and similarity to existing glyphs, respectively;

(ii) Changing the genetic representation of each stencil to enable the encoding of Bézier curves, and this way provide more expressiveness to the stencil;

(iii) Adopting an approach similar to the one developed by [12, 13], allowing the users to design fitness functions through a responsive user interface, thus allowing them to express their design intentions at the meta-level. The user-designed fitness functions are based on features presented by the stencils, namely behavioural features, related to how each stencil performs in drawing glyphs for the target characters, and morphology features, related to the structure and components of each stencil;

(iv) Based on previous experimental results, as the evolutionary process unfolds, the stencils being evolved tend to converge towards an optimum, resulting in visually similar stencils. The lack of diversity problem is not new in the domain of arts and design, and have been addressed by novelty search algorithms by several authors in robotics [14], art [14, 15] and games [14]. We employ an archive mechanism with hybrid tournament selection [16] that allows us to address this issue, promoting diversity among the stencil being evolved and providing a way to summarise the evolutionary runs.

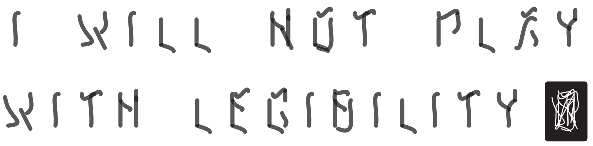


Figure 1: Stencil, and its glyphs, evolved with the presented framework in experiment I.

The experimentation described herein focuses on validating the novel aspects of the approach. We begin by assessing the adequacy of the representation and of the ML-based fitness components by performing experiments on the evolution of stencils that are compact, composed of lines with continuity between them, and that produce glyphs that are recognisable and similar to existing typefaces. Then we test the user interface by performing and analysing runs with user-defined fitness functions. The analysis of the experimental results aims at two core aspects: the ability of the GA to optimise fitness and the ability of the evolved stencils to (i) capture the design preferences expressed by the users and (ii) meet their expectations. Finally, we assess the ability of the novelty search mechanism to generate diverse stencils in a single evolutionary run and the adequacy of the archive, produced during the process, to summarise the results that are then presented to the user.

Overall, the experimental results (see Figure 1) show the adequacy of the proposed approach, demonstrating how EC and ML may address challenges in type design and expand the tools for the optimisation of the design process. Additionally, they also show how meta-level interactive evolution [12, 13] allows the expression of user intentions and goals without imposing a burden to the user, and how novelty search increases the diversity of feasible solutions.

The remainder of this paper is structured as follows. First, we overview the proposed framework. Then, we conduct experiments on the framework, describing the experimental setup and analysing the experimental results. Finally, we present conclusions and directions for future work.

2. Approach

Similarly to our previous work [11], the presented approach is based on the idea of a stencil capable of generating every letter of the alphabet in a coherent manner. In 1876, the American engineer Joseph A. David developed the *Plaque Découpée Universelle* (see Figure 2). This stencil consists of a grid of lines that enables the construction of letters, numbers, punctuation, accents, etc. [17]. The seven-segment display, invented a few decades later, employs a similar approach as the PDU by switching on and off its segments in different combinations in order to represent figures and letters.

The design of typefaces typically involves the creation of modular parts that are then combined by the designer to form different glyphs. By careful looking at the glyphs of

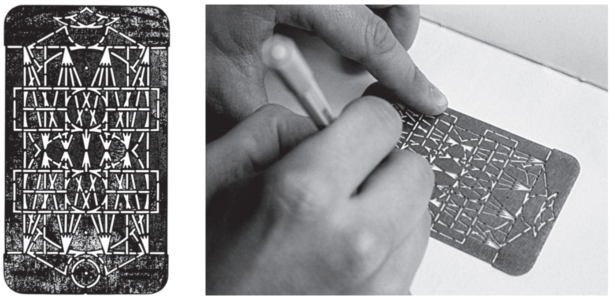


Figure 2: *Plaque Découpée Universelle*, Joseph A. David, 1876.

a given typeface, one may understand their anatomy [18], i.e. the reuse of smaller parts among glyphs. This sharing of parts between glyphs is fundamental to provide them visual coherence. Similarly, the elements (e.g. lines) that construct a stencil can work as a unifying grid that also provides coherence to glyphs created with those elements.

We present a system that employs a Genetic Algorithm (GA) to evolve type stencils (see, e.g., [19] for more details about GAs). The system shares common traits of the works presented in [11] and uses a feasible-unfeasible strategy presented in [16]. We also use an archive to save the evolved individuals based on a similarity criterion [16, 20, 21].

The system is schematically overviewed in Figure 3 and behaves as follows. The evolutionary process begins with the initialisation of the population with randomly created stencils. The fitness of each stencil is calculated according to a fitness function designed by the user. With the stencils evaluated, a new generation of stencils is created using an elitism strategy, i.e. a preset number of fittest stencils proceed unchanged. This step has no effect on the first generation. After the population is evaluated, stencils above a preset threshold are considered feasible stencils. The feasible stencils are compared with the stencils on the archive and if they are dissimilar from the existing stencils they are added to the archive. A stop criterion is tested to determine whether evolution proceeds or stops. If evolution continues, the system determines based on the number of feasible stencils whether novelty search is performed or not, determining how stencils are selected as parents. If novelty search is not performed, stencils are selected by tournament based on their fitness. If novelty search is performed the tournament is based on the fitness and novelty of the individuals of the population and the individuals that are on the archive. Variation operators, *i.e.* crossover and mutation, are applied to the stencils selected as parents to generate offspring stencils. The offspring stencils are evaluated and a new generation is again formed. The whole evolutionary process is repeated until the stop criterion is satisfied. The following subsections detail some mechanisms of the system.

2.1. Representation. Each stencil being evolved consists in a composition of line segments and Bézier curves with varying thicknesses. Therefore, each gene in the genotype of each stencil encodes one line segment or curve in a two-dimensional space.

We implemented an encoding that enables the representation of line segments and Bézier curves. Each gene consists in a 9-tuple with the coordinates of the two endpoints, the angles of the two control points, the lengths of the two control points, and the thickness value. Genes with angle and/or length of the control points set to zero represent straight lines. Figure 4 illustrates the different attributes encoded in each gene: (X1, Y1, X2, Y2, A1, A2, L1, L2, T). The position of the endpoints is constrained by a square grid with a preset density. Also, note that the number of lines may vary from stencil to stencil.

The mapping mechanism that expresses each genotype into its phenotype consists in the drawing of black lines encoded in the genotype on a white canvas. However, the mapping process of the stencils being evolved is not direct. We need one mapping for each character we want to draw with the stencil. This way, we developed a mapping mechanism based on binary masks that define how a given stencil is used to draw a given glyph. When we say *how* we mean which lines are used. This mechanism is illustrated in Figure 5. In what concerns representation, a stencil-based approach may hinder the evolutionary process, because the genetic algorithm has to find a structure of lines (stencil) that is capable of drawing any letter. One could say that we are dealing with a compression problem, i.e. compressing all letters into a stencil. Nonetheless, we believe the "compression" nature behind this stencil-based approach may promote coherence and unity among the resulting glyphs. Furthermore, it helps to understand the anatomy of glyphs and how they share their components/parts. Also, this representation enables us to use an evolved stencil to draw more visual elements other than letters, e.g. signage or symbols, that would be coherent and have the same style as the letters encoded in that stencil.

2.2. Variation. For the initial population the stencils are created at random. We perform variation operations on the stencils using crossover and mutation. The crossover operation consists in the exchange of lines between two stencils. The crossover operator can be summarised to the following steps: (i) select a random rectangular area of the grid; (ii) determine for both parents the lines whose middle points are inside the random rectangular area; and (iii) exchange those lines between the parents. This crossover may be asymmetric as the number of genes it moves from the individual A to individual B may be different from the number of genes it moves from B to A. This results in stencils with a different number of elements in comparison with their parents.

The mutation of a stencil consists in the random modification of genes (lines) of its genotype and comprises three procedures: deletion, modification, and insertion. Each mutation procedure can occur independently with preset probabilities. The deletion procedure selects a line at random and removes it from the stencil. The modification procedure changes one or more lines of the stencil by performing one of the following options, each with a preset probability: (i) moving one of the endpoints by the minimum translation in the grid in one of the eight possible directions; (ii) varying

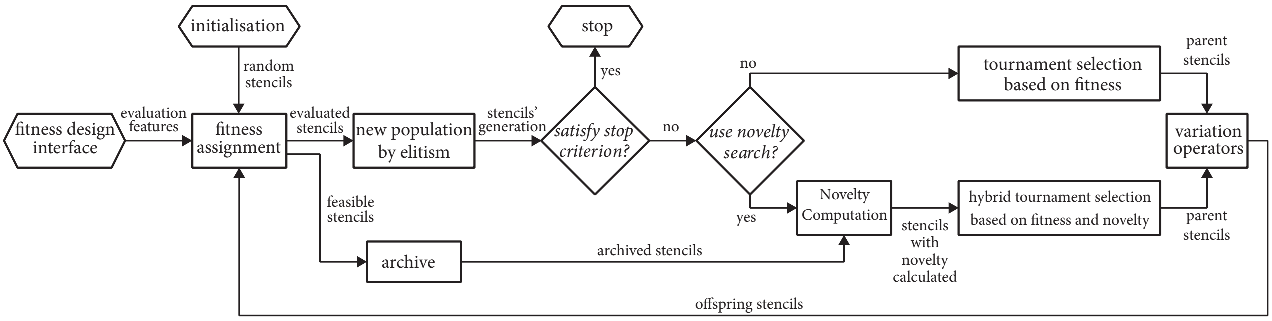


Figure 3: Framework overview.

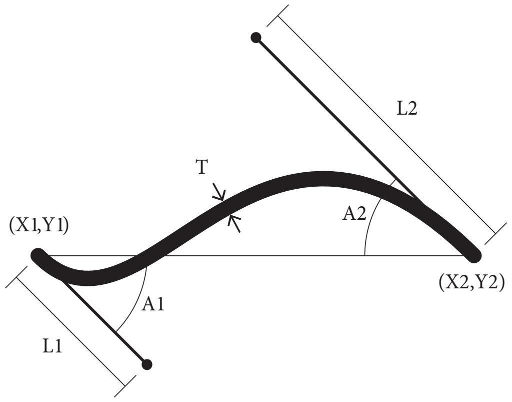


Figure 4: Line encoding.

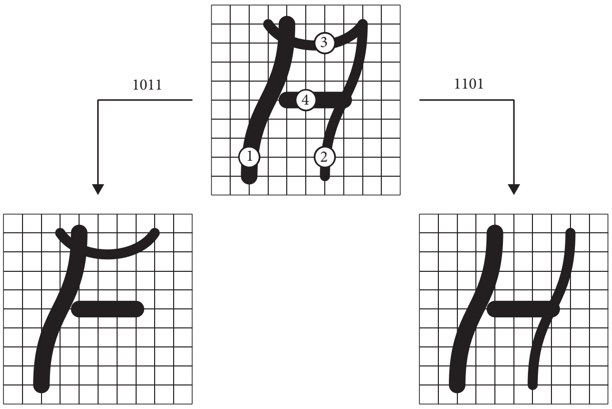


Figure 5: Mapping mechanism expressing the genotype of a stencil (top) into two glyphs (bottom). Binary masks are used to indicate which lines of the stencil should be used to draw the glyphs.

the angle of one of the control points; (iii) varying the length of one of the control points; or (iv) varying the thickness value. Finally, the insertion procedure inserts a new randomly generated line into the stencil. The deletion and insertion procedures cause the variation of the number of lines, enabling the evolution of stencils with different size. Either variation operators preserve the validity of the stencils. A stencil is considered valid if: (i) all its lines are different; (ii) all lines are located inside the limits of the grid; (iii) the number of lines remains within a preset range; (iv) no line has null length; and (v) no line contains another one.

2.3. Evaluation. Based on the work of Romero et al. [22] and previous approaches [9, 20, 23, 24], we adopt an automatic fitness assignment scheme to evaluate the individuals, i.e. stencils, and this way autonomously guide the evolutionary process.

The evaluation of each stencil consists in the computation of (i) behaviour features, related to how the stencil performs in drawing glyphs for the target characters, and (ii) morphology features, related to its structure and components. We conceived the fitness assignment to enable any combination of features to be pursued by the evolutionary process. Furthermore, in a combination of features, each feature can have more or less importance (weight) in comparison to the other features. As a result, the fitness function consists in a weighted product:

$$fitness\ (ind) = \prod_{i}^{n} \left(satisfaction_i\,(ind) * w_i + 1 - w_i \right), \tag{1}$$

$$satisfaction\,(ind) = 1 - \left| featurevalue_i\,(ind) - t_i \right|, \tag{2}$$

with *wi*, *vi*, *ti* *[0,1]*, where w_i is the weight of the *ith* feature, t_i is the desired target value for the feature *i*, and $featurevalue_i$ is the value measured corresponding to the feature *i*. The evolutionary process aims at maximising the value of Equation (1), whose theoretical optimum corresponds to a stencil that simultaneously matches all features to their target values according to Equation (2). All the features and weights are normalised to the [0..1] domain. As such, maximising or minimising a given feature consists in setting its target value to 1 or 0, respectively.

The following subsections detail the behaviour and morphology features of the stencils.

2.3.1. Behaviour. One of the preconditions of the stencils evolved with the proposed framework is their ability to produce glyphs that are legible, *i.e.* stencils should be able

Table 1: Behaviour features.

feature	description
recognisability	character recognition using a Convolutional Neural Network (confidence value of the classifier in recognising the potential glyph as the target character)
similarity	visual similarity to existing glyphs using Self-Organising Maps (RMSE pixel-by-pixel similarity between the potential glyph and the most similar neuron in the Self-Organising Map of the target character)

to express images that are recognised as characters. The framework evaluates this ability by measuring features based on the glyphs that each stencil is able to express. Table 1 presents an overview of these features, which we refer to as behaviour features.

The expression of each stencil into glyphs takes a couple of steps. As explained before, each stencil has several lines that can be activated to draw glyphs. First, the system chooses a character for which the stencil has to produce glyphs. Next, among the lines that compose the stencil, we search which mask of active lines better expresses glyphs for the chosen character. This way, each mask stores the best use, or configuration, of the stencil found during the evaluation process to draw a given character.

We use a hill-climbing algorithm to perform the search for the best configuration of the stencil being evaluated for each target character. This way, the evolutionary process includes a nested search to find optimal configurations for each stencil. The search starts with all the lines deactivated, activating one per step. At each step, all newly generated configurations are evaluated as a glyph for the target character, *i.e.*, how the expressed glyph matches the target similarity and recognisability values. The search stops when no improvement in the evaluation is achieved, storing the best mask and evaluation value of the resulting glyph. This process is repeated for all the target characters. In a typical evolution, all target characters are equally considered, i.e. the stencils being evolved should be able to draw glyphs for all of them. However, the user can specify different importance levels for the target characters. This enables the user, for example, to evolve stencils that can express glyphs for a subset of characters, or to improve the legibility of some glyphs of an evolving stencil.

Recognisability. We use a Convolutional Neural Network (CNN) classifier to calculate how much a glyph is recognised as a given character. CNNs are a type of Deep Neural Networks (DNNs) that have been used successfully in image classification and recognition tasks [25, 26]. The main characteristic of a CNN is the usage of convolutional and pooling layers, which provide feature extraction and dimensionality reduction in training [27]. Each layer can be seen as a filter from which features are extracted and learnt.

The architecture of the CNN is based on the *Lenet*-5 network for digits recognition [28] but trained as multi-class supervised classifier for the 26 capital letters of the Roman alphabet. Our approach must perform several evaluations of several glyphs per generation, therefore the chosen network was a trade-off between computational power and efficiency. The classifier is trained on the 32-by-32 pixel representation of the typefaces served by Google Fonts. Besides the typefaces, we added a negative class represented with random images generated by our approach that do not resemble any characters, yielding a total of 27 classes. The value of the recognisability feature of a stencil configuration is the output of the classifier, indicating its confidence in recognising the input image of the configuration as its target character. An output of 1 indicates total confidence while an output of 0 indicates the opposite. Note that the Machine Learning model used is based on data available and off-the-shelf architectures with its inherited limitations and exploits [29, 30]. Furthermore, the models employed are not able to harness the full potential of the human-like visual system. With this in mind, this feature is used to evaluate the legibility, recognisability and readability of the input images.

Similarity. An array of Self-Organising Maps (SOMs), one for each target character, is used to calculate the similarity feature. The SOM [31, 32] is among the most well-known unsupervised learning and clustering approaches. The architecture of the SOM consists in a feed-forward neural network that reduces information while preserving the most important topological relationships of the data elements. This enables the calculation of the visual similarity of the glyphs expressed by each stencil with existing glyphs.

Each SOM is constructed of 64 neurons and is trained with 32-by-32 pixels images of several glyphs of the corresponding target character. The glyphs used for training were gathered from the typefaces of the Google Fonts platform.

To calculate the similarity feature of a stencil configuration (glyph), an image representation of it is created with the same size as the SOM neurons and then compared with each neuron of the target character SOM. In this comparison, the similarity between the image representation of the stencil configuration and each neuron in the SOM is calculated using the Root Mean Square Error (RMSE), which measures how close, or far, the candidate glyph is to a reference glyph (neuron) on a pixel-by-pixel basis. The value of the similarity feature considers the distance between the glyph expressed by the stencil and the most similar neuron in the SOM, also called the best matching unit, and is given by: $1 - \text{normalised}_{\text{RMSE}}$.

2.3.2. Morphology. In addition to the behaviour features, the framework considers a series of other features related to the structure and components of the stencil. Table 2 presents an overview of these features, which we refer to as morphology features.

By adjusting the morphology features, one can promote the evolution of stencils that exhibit particular structural

Table 2: Morphology features.

feature	description
size	number of stencil lines (normalised to the [0..1] range according to a preset range)
coverage	rectangular area occupied by the stencil lines (normalised to the [0..1] range according to the grid area)
continuity	percentage of stencil lines that share endpoints with other line
intersection	percentage of stencil lines that intersect other line
parallelism	percentage of stencil lines that are parallel to other line
horizontal symmetry	similarity between the top half of the stencil and the bottom half mirrored vertically (calculated using the RMSE between the top and bottom half)
vertical symmetry	similarity between the left half of the stencil and the right half mirrored horizontally (calculated using the RMSE between the left and right half)
curves	percentage of stencil lines that are curves
symmetric curves	percentage of stencil curves that are symmetric in relation to the line that (i) is perpendicular to line segment S and (ii) intersects the middle point of S; where S is the line segment that connects the two end points of the curved line.
length	average length of the stencil lines (normalised to the [0..1] range)
length diversity	standard deviation of the lengths of the stencil lines (normalised to the [0..1] range)
thickness	average thickness of the stencil lines (normalised to the [0..1] range)
thickness diversity	standard deviation of the thicknesses for the stencil lines

characteristics. The possibilities are vast. For instance, one can configure the morphology features of the fitness function to reward stencils with only curved lines that intersect little, or horizontally symmetric stencils with only straight lines with great continuity, or stencils with long curved lines that intersect little.

The combination of morphology and behaviour features enables the evolution of stencils that match particular visual characteristics while ensuring the legibility of the glyphs. This way, the user can explore different compromises between the legibility and the style provided by the generated stencils.

2.4. Archive. Similar to the work done in [16], the archive is used to evaluate our solutions during the evolutionary process and prevents the algorithm from searching areas of the search space that were already visited. The archive should hold the set of stencils found to date by the evolutionary process. The size of the archive shows how the algorithm can generate diversified stencils.

The archive comes into play after the fitness assignment. At this stage, a candidate stencil has its fitness assigned, and it has to meet two requirements in order to be added to the archive: (i) its fitness must be greater than or equal to an adequacy threshold f_{min}; (ii) it needs to surpass a dissimilarity threshold when compared to those that already belong to the archive.

This process is performed by computing the average dissimilarity between the candidate and a set of k-nearest neighbours. When the average dissimilarity is above a predefined dissimilarity threshold, $dissim_{min}$, the individual is added to the archive. In this approach, we evaluate the stencils based on their expression as glyphs, i. e., the stencils are analysed in the form of images of their glyphs. As in [16], the dissimilarity metric for an image i is computed as:

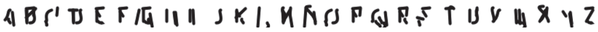


Figure 6: A stencil's expression rendered to a single image, which is used to compute the similarity between stencils for the archive algorithm.

$$dissim(i) = \frac{1}{max_{arch}} \sum_{j}^{max_{arch}} d(i,j) \qquad (3)$$

Where max_{arch} is a predefined parameter for the number of most similar images to consider when comparing with image i and $d(i,j)$ is a distance metric. The distance metric measures how different two images are. There are two exceptions to the application of this measure: (i) if the archive is empty then the first stencil that has a fitness above the f_{min} is added; and (ii) if the number of entries on the archive is less than max_{arch} we use the number of existing entries instead of max_{arch}.

In order to evaluate the similarity between stencils, we resort to an image similarity metric applied to the stencil's behaviour, i.e. the image output of the configurations for each letter. One image is created containing several letters concatenated to form a "banner" image as shown in Figure 6.

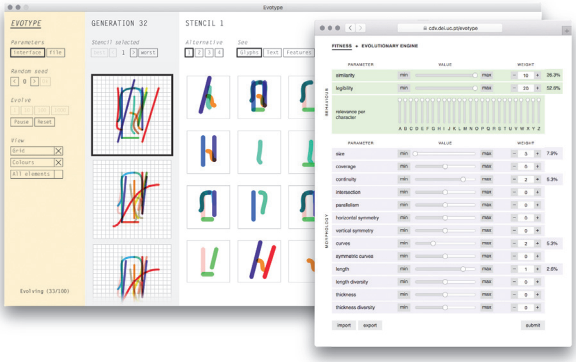
The banner is used to evaluate the similarity among the several candidate stencils and the archived ones. We use a similarity metric the RMSE between the pixels. It is out of the scope of this work to explore several dissimilarity metrics. Since we are processing a considerable number of images per generation, we resorted to RMSE for its fast calculation. To the interested reader, we suggest consulting the works by [33, 34] for more detail on similarity and dissimilarity metrics. When a stencil is added to the archive, it counts as a feasible solution.

The mechanism that selects feasible solutions is important to shape how evolution will proceed, depending on the results obtained in a given generation. We introduce the novelty approach presented in [16], a customised selection mechanism that can switch between a fitness-based strategy and a hybrid mechanism that considers both fitness and novelty. As depicted in Figure 3 it can switch between fitness and hybrid according to the following decision rule: if the number of feasible solutions of the current generation is lower than a threshold T_{min} change to fitness guided evolution; or if the number of feasible solutions of the current generation is above a threshold T_{max} change to hybrid mechanism. In fitness guided evolution, the tournament selection is based on the fitness values of the candidate solutions, as in a standard Evolutionary Algorithm (EA). If hybrid is chosen, it is necessary to compute the novelty of each selected individual, and perform a Pareto-based tournament selection, using the novelty and fitness of each selected individual as two different objectives to maximise. The novelty computation process is inspired by Lehman and Stanley's work [35], with one small change: the *k* most similar images are considered from the set of the selected individuals and the archive, instead of considering the whole population and the archive. At this stage, each selected individual has a fitness and novelty value, and there is the need to determine the winner of the tournament. This process is inspired by multi-objective EAs, namely the Pareto-based approaches, which select the best individuals based on their dominance or non-dominance when compared to other individuals. In this work, the hybrid tournament selection determines the non-dominant solutions by comparing, among the selected individuals, both fitness and novelty. After the set of non-dominant individuals are computed, we have the so-called Pareto front. Using the hybrid mechanism, the tournament winner is selected by randomly retrieving one of the solutions of the Pareto front.

2.5. Implementation. The proposed framework is implemented in two modules: (i) the evolutionary system and (ii) the fitness function design interface. The first module, the evolutionary system, operates the GA and the fitness assignment scheme that automatically guides it. The second module, the fitness function design interface, enables the user to adjust parameters of the fitness assignment and other inner workings of the first module. Figure 7 shows a screenshot of the evolutionary system (left) and the fitness function design interface (right).

The fitness function design interface communicates with the evolutionary system through a JSON file. Technically, one could manually adjust the parameters stored in that file and this way use the evolutionary system alone to evolve stencils. However, this would hinder the design of fitness functions and the configuration of the evolutionary process.

A typical use of the framework could be initiated as follows. The user launches the evolutionary system and selects the source of the setup parameters: (i) from a setup file or (ii) from the fitness function design interface. When the first source is selected, the user imports a setup file stored in the computer, *e.g.*, a setup file previously exported using the fitness function design interface. This approach is useful to test a series of experimental setups. When the second source is selected, the evolutionary system activates a mode in which it listens for new parameters coming from the fitness function design interface. In this approach, the user uses the fitness function design interface to adjust the setup parameters, which are directly sent to the evolutionary system. This enables the user, for example, to modify the fitness function during the evolutionary process. After selecting the source of the setup parameters, the user is in position to evolve stencils. In the following subsections, we overview some of the key functionalities of the two modules.



PARAMETER	VALUE		WEIGHT	
similarity	min	max	10	26.3%
legibility	min	max	20	52.6%
size	min	max	3	7.9%
coverage	min	max	0	
continuity	min	max	2	5.3%
intersection	min	max	0	
parallelism	min	max	0	
horizontal symmetry	min	max	0	
vertical symmetry	min	max	0	
curves	min	max	2	5.3%
symmetric curves	min	max	0	
length	min	max	1	2.6%
length diversity	min	max	0	
thickness	min	max	0	
thickness diversity	min	max	0	

Figure 7: Screenshot of the framework, consisting of the evolutionary system (back) and the fitness function design interface (front). A demo video can be seen at http://cdv.dei.uc.pt/2018/complexity/evotype.mov.

2.5.1. Evolutionary System. The evolutionary system module provides the necessary means to evolve, browse, test, and export type stencils. After selecting the setup parameters, one can command the evolution of stencils by setting the random seed and instructing the system to generate a given number of generations. During evolution, it is possible to browse throughout the current generation of stencils, which are arranged vertically by descending order of fitness. The system features a mode that renders the elements of each stencil using different colours, either when previewing the entire stencil or the glyphs produced with it. The purpose of this mode is to visualise the reuse of elements of the stencil between the different glyphs. The user can select each stencil to (i) test it by typing a couple of words with the glyphs produced with it; (ii) visualise its features, or measurements, that are being considered by the fitness function; and (iii) export it to file, enabling further refinements and its utilisation outside the framework.

2.5.2. Fitness Function Design Interface. The module of the fitness function design interface enables one to design fitness functions to automatically guide the evolution of type stencils. The interface empowers the user by enabling him/her to express preferences through the specification of properties that he/she intends to observe in the evolved stencils. Although the main goal of the interface is the configuration of the fitness function, it also enables the adjustment of several parameters of the evolutionary process, *e.g.* population size, elite size, tournament size, crossover rate, mutation rate, grid size, phenotype size, and other parameters related to the novelty search approach, including the minimum fitness for a stencil to be considered feasible and the minimum dissimilarity to be added to the archive.

The fitness function design interface consists of a web page with multiple sliders and buttons that enable one to adjust evaluation and evolution parameters in a high-level way. One could say this module acts as an interactive facilitator of parameterisation of the first one, the evolutionary system, abstracting the user from the inner workings of the framework. This approach enables the user to submit parameters to the evolutionary system at any time, as already explained, and this way develop or change his/her preferences throughout the generations. In addition to submitting the current parameters to the evolutionary system, the user can also export the parameters to file and import them later. The decision of implementing this module as a web page is related to our short-term goal of converting the framework into a web application. This would enable anyone to use the framework.

Based on the two levels of evaluation identified at the beginning of this section, the fitness parameters presented in the interface were arranged into two groups that are visually distinguished using different colours. The top group of parameters is related to the behaviour of the stencil, while the bottom group is related to its morphology. The interface employs tooltips to enable the user to understand the different components of the interface, *e.g.* the semantics associated with each feature and how it is calculated. When the user hovers the cursor over a component, a tooltip appears displaying information about it.

For each parameter, the user may set (i) the value that should be matched by the stencils being evolved and (ii) a weight that indicates the importance of that parameter in the fitness function. The only exception is the last parameter of the behaviour group, which presents an array of vertical sliders to specify the relevance of each character the evolved stencils should be able to draw glyphs for. Changing the weight of one parameter results in having more or less impact in comparison to the other parameters. In order to make the adjustment of weights easier to understand, we adopted an approach where the user indicates each weight by adding or subtracting units to the weight. Nonetheless, one can also set the weight to a specific floating value. For instance, one parameter with a weight of 3 would have an importance three times greater than a parameter with a weight of 1. Following this reasoning, setting the weight of one parameter to 0 means that it will be ignored. One advantage of this approach is the precision it provides to the user when adjusting each weight, in comparison to other approaches that employ, for example, sliders. The final weight of each parameter, considering the other weights, is displayed on the right side. This information helps the user understanding the overall impact of each individual parameter in the fitness function. Also, by only displaying the final weights greater than zero, we are able to visually highlight the parameters that are being considered.

3. Experimentation

We conduct four experiments on the proposed framework with different goals in mind. In the first experiment, we study the adequacy of the hybrid fitness function (similarity and recognisability) for guiding the evolutionary process. In the second experiment, we analyse how the design of fitness functions influences the outcomes of the system and if, and to what extent, it is able to convey the preferences of the user. In the third experiment, we investigate the impact of novelty search on the evolutionary convergence and on the diversity of stencils evolved. In the last experiment, we explore the creative possibilities provided by the outputs of the presented framework by using one evolved stencil in a design project.

In this work, we evolve stencils to draw glyphs for the uppercase letters of the Roman alphabet. The base experimental parameters are summarised in Table 3.

3.1. Experiment I - Hardwired Fitness Functions.

In this section, we analyse the ability of the approach to evolve stencils with hardwired fitness functions. In [11] we validated that the evolutionary engine by performing experiments that used and hardwired fitness function resorting to RMSE for a predetermined target typeface. The results have shown that the evolutionary algorithm is able to evolve stencils that expressed visually coherent glyphs. However, in the first set of experiments, the evolutionary algorithm generated stencils which maximised the number of elements and some presented several gaps. In the second set of experiments, we redefined the fitness function to control the number of elements and minimize gaps. The approach was able to generate simpler stencils able to produce glyphs similar to the targets using lesser elements, promoting the reuse of stencil's elements for multiple glyphs. An overall observation is that in order to promote a specific behaviour we have to redefine the hardwired fitness function, which is a sensible and time-consuming process.

In the experiments of [11] the only behaviour feature used to guide the fitness was the similarity feature. It was based on the pixel-based RMSE between a glyph expression to a predefined target typeface glyph. In order to promote more flexibility in the solutions, we use a SOM for the calculation of the similarity feature. The SOM organises and reduces the instance space. We use RMSE to compute the similarity of a candidate glyph expression with an expression of the closest SOM neuron. The SOM trained with several typefaces provides different targets to explore while reducing the number of targets to be evaluated. This allows for a more flexible evaluation of similarity. Nevertheless, there is a possibility of the approach exploring the activations of different SOM neurons belonging to different glyphs. We also introduce the concept of recognisability, performed by a CNN. The idea of using the CNN evaluation as part of the concept of recognisability is to promote the generation of stencils which retain recognisable characteristics of existing glyphs.

In Table 4 we present three fitness functions defined with different recognisability and similarity weights. The size and continuity features were maintained from the experiments

Table 3: Experimental Parameters.

parameter	value
generations	1000
population size	100
elite size	1
selection	Tournament
tournament size	3
rate crossover	0.5
max genes	40
rate deletion	0.05
rate insertion	0.05
rate modification	1 / genotype size
grid size	20 x 20
min length permitted (value relative to grid size)	0.15
control points angles permitted (rotation angles in degrees and relative to the line segment that connects the endpoints)	$[-90, -45, 0, 45, 90]$
control points lengths permitted (values relative to the distance between the endpoints)	$[0.25, 0.5, 0.75]$
thickness values permitted (values relative to the phenotype size)	$[0.125]$
phenotype size	32 x 32
hill-climbers	1

Table 4: Fitness functions designed and tested in experiment I.

fitness function	feature	target value	weight
fitRec	recognisability	1	100
	similarity	-	0
	size	0	2
	continuity	1	2
fitSim	recognisability	0	0
	similarity	1	100
	size	0	2
	continuity	1	2
fitHybrid	recognisability	1	67
	similarity	1	33
	size	0	2
	continuity	1	2

in [11]. In all the experiments we track the response values of each feature that compose the final fitness function for analysis purposes, even if the weight is set 0, i.e. it does not participate in the calculation of the fitness.

In *fitRec*, the evolutionary process is mainly guided by the recognisability feature, i.e., based on the activation of the CNN for each stencil's glyph expression. In Figure 8, on the top left plot we can observe the behaviour of the evolutionary algorithm using the *fitRec* fitness function, showing that we are able to guide evolution and optimise the stencils' fitness. It starts with a relatively low fitness value but rapidly converges to high fitness values in a few generations. If we analyse the progression of the values of the other components, we can observe that the similarity is not affected by the progression of the recognisability. The size feature in the first generations tends to rapidly increase, meaning that elements are being removed from the stencil. When the fitness stabilises, the size feature increases, expressing the highest value amongst the tested functions. It means that in the end, it has fewer elements than the other fitness functions. Regarding continuity, it consistently rises along the generations. Based on the values at the end of the evolutionary process, it seems to create disconnected stencils when compared with the other fitness results.

The *fitSim* function uses the similarity feature to guide fitness. As shown in Figure 8, the evolutionary algorithm is able to optimise the fitness function, although it does not reach the maximum theoretical value. The recognisability values tend to increase with the increase of the similarity feature values. The size component consistently increases, suggesting that the best stencil tends to remove elements along the generations. When compared to the others, *fitSim* reaches to the highest number of elements used by the stencil. The high values of the continuity feature show that it tends to create a connected stencil.

The *fitHybrid* is a fitness function that combines both the similarity and the recognisability to guide evolution. The values of Table 4 show that more weight was given towards the recognisability. The idea is to have stencils able to produce expressive and functional typefaces, exploring different compromises between the expressiveness and the legibility of the glyphs while maintaining coherence. Once again, we

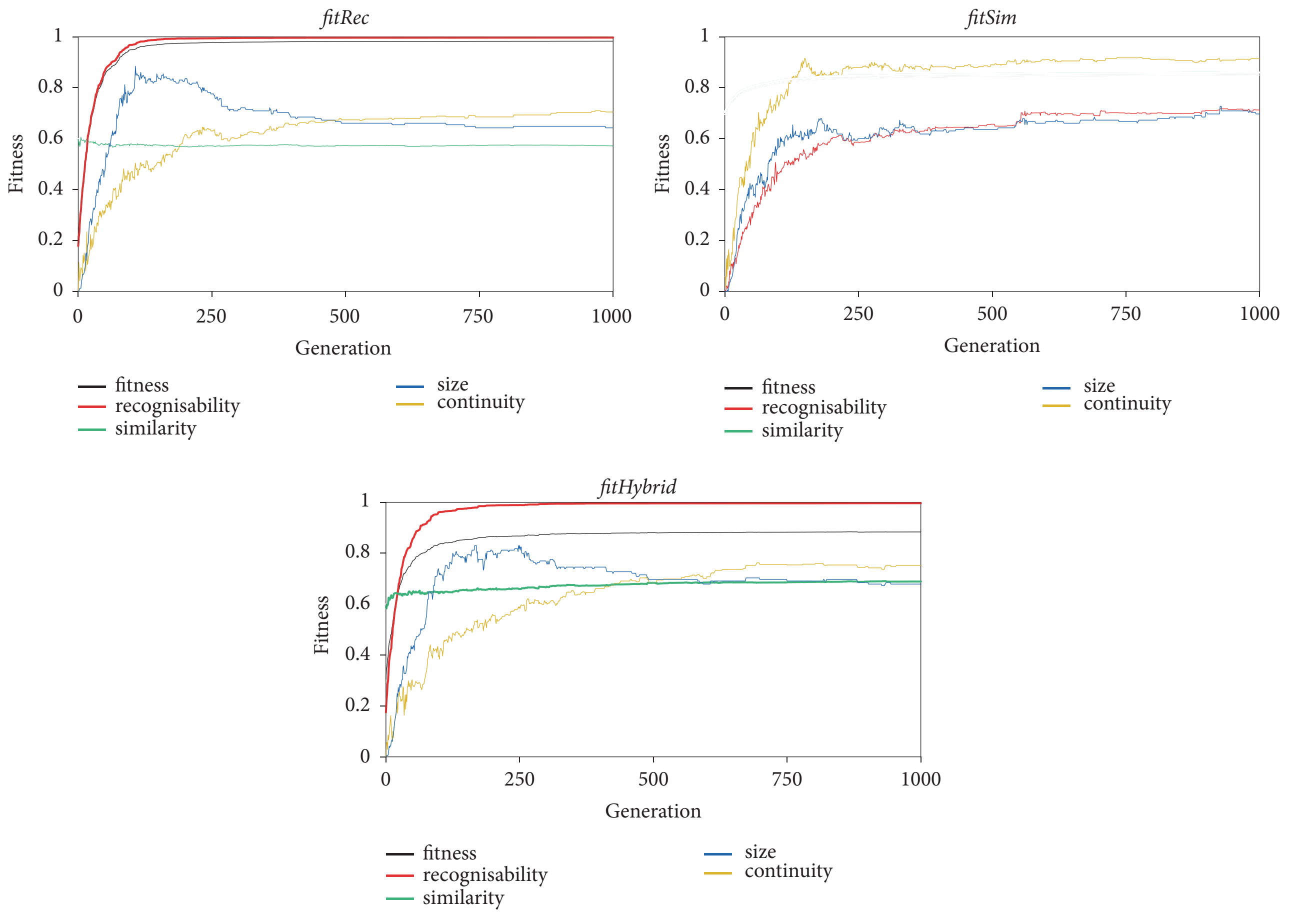


Figure 8: Progression of the fitness and features' values of the fittest stencil when fitness functions *fitRec* (top left plot), *fitSim* (top right plot), and *fitHybrid* (bottom plot) guide the evolutionary process. The visualised results are the average of 30 runs.

are able to guide the evolutionary process and optimise the fitness function. In terms of fitness, it is possible to observe that it maintains a certain stable level of similarity and that the recognisability contributes more to the fitness increase. In terms of the other two features, size and continuity, we can say that we get the good from both fitness features, i.e., a low number of elements and a more connected stencil.

In Figure 9 we can observe generated stencils for the fitness function used in this first experiment. It is noticeable the difference between them at the visual level. The results also cope with the analysis in terms of fitness components. *fitSim* tends to generate stencils with more and connected elements. Although the SOM gives us more flexibility than the RMSE target approach of [11], in this approach it generates stencils that fill the space of the target neurons of the SOM. The flexibility comes with a trade-off, some of the glyphs generated by the stencil appear to focus on different SOM neurons, resulting in different glyphs expressions. *fitRec* uses fewer elements but the used elements are more disconnected and dispersed. However, we see some random features around the generated glyphs that can be seen as exploits of the classifier guiding the evolution. We are aware of the propensity of the evolutionary algorithms to find shortcuts and exploit weaknesses on fitness assignment schemes that use ML approaches [28–30]. In *fitHybrid* we combined the recognisability with the similarity to prevent the approach to guide the evolution to recognisable but atypical glyphs. *fitHybrid* tends to generate stencils with a small number of elements that are connected generating glyphs that are simpler, distinguishable elements, demonstrating variability while maintaining coherence. Overall, we consider that using the fitness functions defined for this experiment we are still able to evolve stencils that generate coherent glyphs.

In the previous set of experiments, we used a stricter evaluation based on a target font [11]. The algorithm converged to a structural representation of that font, i.e. a stencil to draw it, showing that the approach can arguably generate a compressed representation of a font. In the experiments presented in this section, we observe a similar behaviour. This enforces the idea that the representation is adequate to the task at hand.

The results show that we can evolve recognisable and legible fonts, but this is not enough for them to be aesthetically appealing. Performing more generations could marginally augment the aesthetics of the results but would not lead us to the high-quality solutions of a commercial font. This fact leads us to two different hypotheses not mutually exclusive about type design. When a type designer creates a font, it does

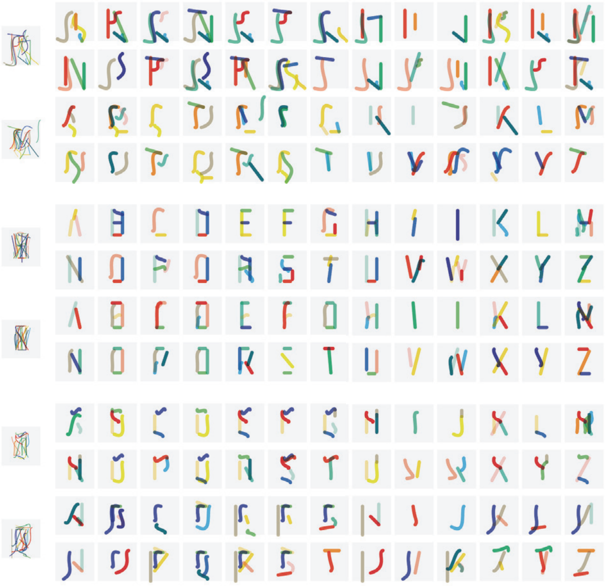


Figure 9: Typical results evolved in different runs of *fitRec* (top group), *fitSim* (middle group) and *fitHybrid* (bottom group). To better identify each element of the stencils (left) in the corresponding glyphs (right), a random colour is used for each element.

not look exclusively into its functionality [31–33]. The visual features and ML approaches in use are not able to harness the potential of the human-like visual system to guarantee that if something is legible from the machine point of view is legible for the human and that the factors that lead to an increase or decrease in terms of legibility, readability and recognisability to a human are the same for the machine and vice-versa. Assuming that this is the case, if we use more complex visual models, trained to recognise other types of artefacts and, as such, subject to the tasks that a human is subject to deal in a daily basis it could contribute to enhancing the results.

Based on overall results and discussion, we consider that experimenting with fitness function design in a semi-automatic way can be beneficial. This path is explored in the next set of experiments.

3.2. Experiment II – Designing Fitness Functions. In this experiment, we analyse how the design of fitness functions influence the outcome of the framework and if, and to what extent, they are able to convey the specified preferences. We focus on the morphology features because these are likely to be perceived visually on the stencil as well as on the resulting glyphs.

Using the experimental setup tested in experiment I as a base, we add other features to the fitness function. We design and test 4 more fitness functions. Each one consists in the base fitness function (*fitHybrid*) combined with one, or two, more morphology feature(s). The features added to *fitHybrid*, along with a name for the resulting fitness function, are listed in Table 5.

Table 5: Fitness functions designed and tested in experiment II.

fitness function	feature	target value	weight
fitCurves	curves	1	4
fitNoCurves	curves	0	4
fitSymCurves	curves	0.5	4
	symmetric curves	1	4
fitUniformLength	curves	0.5	2
	length	0.66	4

Although many different combinations of features could be tested, we selected some that we believe can be more noticeable in the evolved stencils. Since the encoding of Bézier lines is an iteration of this framework (comparing to our previous work [11]), we also focused this experiment on features related to them.

We designed each fitness function to set the framework to evolve stencils with specific visual characteristics:

(i) *fitCurves* — stencils entirely composed of curves;

(ii) *fitNoCurves* — stencils with no curves;

(iii) *fitSymCurves* — stencils with half of their elements being symmetrical curves;

(iv) *fitUniformLength* — stencils also with half of their elements being curves, and all lines should have a length of two-thirds of the grid size.

Figure 10 summarises the results of this experiment. Per fitness function, we visualise the progression of each feature being evaluated and present one typical stencil evolved using that fitness function. In general, and based on the different runs of each fitness function, the results indicate that: (i) different fitness functions lead to different stencils; (ii) different runs with the same fitness function converge to different stencils, thus providing diverse stencils; and (iii) the four fitness functions designed are able to guide evolution towards stencils with features that match the preferences behind them, sometimes in surprising ways.

The framework frequently finds interesting ways to match the fitness functions, generating unusual glyphs more or less functional. On the other hand, sometimes, the framework generates stencils that, from the type design point of view, may be appealing due to their novelty and aesthetics, but which do not maximise all features of the fitness function. Nevertheless, the results reveal the effectiveness of the approach in exploring possibilities that are consistent with the preferences expressed by the user who designed the fitness function. Looking at each stencil, and their glyphs, in Figure 10, one can see that they exhibit visual properties that are aligned with the features added to each fitness function. For instance, the stencil evolved with *fitCurves* is almost only composed of curves (only one line segment is used); on the other hand, the stencil evolved with *fitNoCurves* is almost only composed of line segments (only three curves are used); the stencil evolved with *fitSymCurves*, in addition to using the same number of curves of line segments, most of the curves used are symmetric (only one curve is asymmetric); and the stencil evolved with *fitUniformLength*, has the same balance of curves and line segments as the previous stencil all

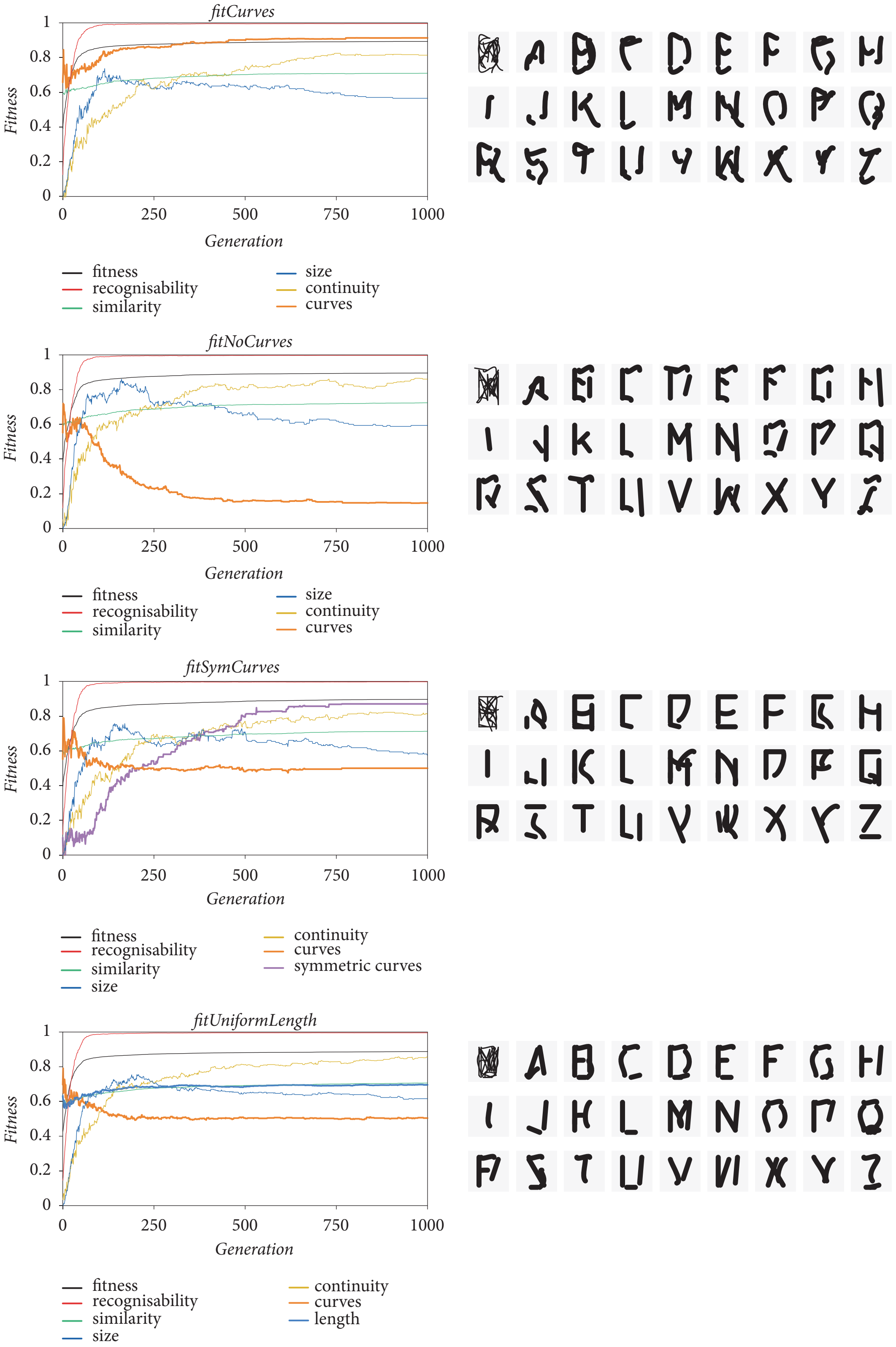


Figure 10: Experimental results when the evolutionary process is guided by *fitCurves*, *fitNoCurves*, *fitSymCurves*, and *fitUniformLength*, in descending order. For each fitness function, one can see the progression of the fitness and features' values of the fittest stencil over the generations (left) and one typical fittest stencil of the last generation (right). The visualised results are the average of 10 runs.

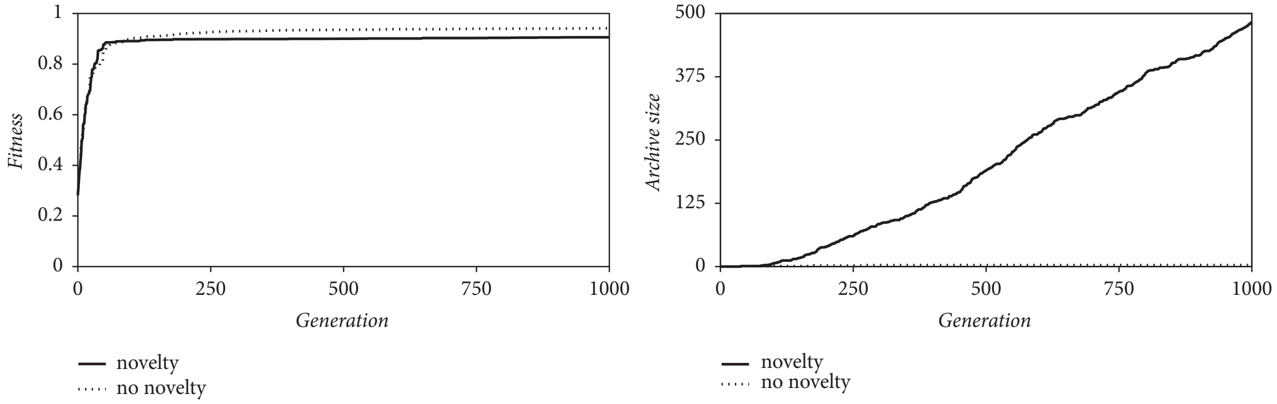


Figure 11: On the left, the progression of the fitness and features' values of the fittest stencil across generations using the *novelty* and *no novelty* setups. On the right, the evolution of the archive size across generations. The visualised results are from a single run.

Table 6: Experimental parameters.

parameter	value
t_{min}	5
t_{max}	15
max_{arch}	5
f_{min}	0.85
Dissimilarity metric	RMSE
$Dissim_{min}$	0.66

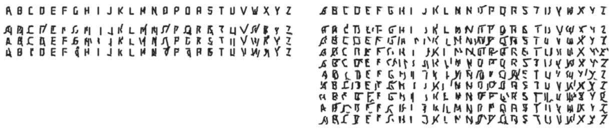


Figure 12: On the left, the archive (sampled) of the expressions from a run of evolving stencils guided by the hybrid mechanism. On the right, the archive (sampled) of a run evolving stencils guided by fitness. The top row instances are the fittest stencil found and the remaining ones above them are the archive for the *no novelty* (on the left) and *novelty* (on the right).

elements have more or less the same length, approximately two thirds the size of the stencil grid.

Overall, evolution was able to optimise all features being tested without compromising the behaviour features, i.e. similarity and recognisability of the glyphs. This should be related to the substantial differences between the high weights used for the behaviour features and the low weights used for the morphology features.

3.3. Experiment III – Novelty Search Mechanisms. In this experiment we assess the ability of the novelty search mechanism to generate diverse stencils in a single evolutionary run and the adequacy of the archive, produced during the process, to summarise the results that are presented to the user. We preserve the experimental setup of *fitHybrid* from the experiment I and used the parameters in Table 6 to be used by the novelty mechanism.

For this experiment, we perform a single run of the following setups: *novelty* - uses the novelty search mechanism; *no novelty* - does not use novelty search, the fitness guided approach. The archive is used on both runs to analyse the impact of the novelty mechanism.

In terms of fitness, in both cases we have a behaviour similar to the one observed in experiment I, it optimizes the fitness function. However, it is noticeable the differences between the two setups in terms of fitness values along the generations. In generation 50 we have the first entry to the archive on both setups, i.e. at least one individual that surpasses the f_{min} value. A few generations after that point, the novelty setup enters in hybrid tournament and we can observe that it rapidly increases up to a certain point, surpassing even the values observed for the fitness guided in the same interval, for a few generations. Around the 100^{th}, generation the *no novelty* setup surpasses the maximum value and continues increasing for a few generations until it stabilizes. We can observe that for the *novelty* setup it continues to slowly increase until the last generation. If we observe in Figure 11 the archive size, we can see clearly that the novelty setup adds much more instances to the archive, suggesting that adds a lot of diversity to the evolutionary process.

Figure 12 shows samples from the archive and the fittest stencils found in *no novelty* and *novelty* setups. We start by analysing the fittest stencils (top left and top right images of Figure 12), observing that they are different from each other which suggests that using novelty impacts the final result of the evolutionary process. Moving towards the analysis of the archives, on the left we show that the *no novelty* setup only adds three stencils to the archive, indicating that for the defined parameterization it does not found any stencil behaviour more dissimilar than the three that we observe in Figure 12. There is some dissimilarity among the stencils but some of the letters remain very similar, e.g., "I", "J", "l" and "Z". We can see that the archive stencils' letters have some resemblance with the fittest stencil's letters. In novelty setup, while analysing the values in the archive size, we see that a lot of different stencils are added to the archive. Due to the high number of entries on the novelty setup archive we employ a filter using RMSE to select the top-10 most dissimilar instances on the archive. We clearly have some diversity amongst the most dissimilar entries. Some of them contain a mixture of clear letters with some letters that from a subjective standpoint do not resemble the corresponding letter (e.g. Figure 12 on the right in the 6th, 9th, 10th rows). Overall, we can see that the evolutionary process explores a larger area of the search space when compared with the *no novelty* setup. Note that since we are only performing a single run, we can say that is possible to create more diversity and generate a more diverse archive of solutions when compared to the traditional fitness guided solution (*no novelty*). The trade-offs are in the extra parameterization that can be difficult to tune, and it could come at the cost of later convergence. Aside from the extra fine tuning, this suffers from the same problem of tuning hardwired functions, since the user is only a spectator on this process once the evolutionary process starts.

3.4. Experiment IV – Applying Evolved Stencils. In this last experiment, stencils evolved with the presented framework are applied in a real design project: an interactive installation integrated in a permanent exhibition dedicated to Portuguese literature that enables visitors of a museum to generate their own portraits made of letters.

The creation of imagery using text is a traditional design task and is nothing but new. The process of drawing with text can be traced back to manuscripts from many centuries ago with illustrations made of handwritten words. Fast forward to the late 1890s, Typewriter Art becomes an art form, with the first piece of known Typewriter Art being documented, an image of a butterfly created by Flora Stacey in 1898 [36]. Later, in 1966, at Bell Laboratories, Kenneth Knowlton and Leon Harmon created the image “Studies in Perception #1”, one of the earliest known examples of ASCII art and probably the first computer nude. To create the image, Knowlton and Harmon scanned a photograph and assigned typographic symbols to the binary numbers according to halftone densities [37].

The interactive installation employs a generative process based on ASCII art to create the portraits. It dynamically changes the typographic weight of each letter to render different shades of grey and this way depict an input image.

The mapping of the darkness of the input image into letters is best accomplished using a typeface with many weights, so a continuous range of shades of grey can be rendered typographically. We believe stencils evolved by the presented framework can play a role here because by using a stencil, each letter can be drawn with as many thickness values as needed. In other words, a continuous range of thickness values can be used to give body to the letters in the portrait and this way enable the representation of different shades.

The video at http://cdv.dei.uc.pt/2018/complexity/evotype.mov shows the interaction with the framework in order to design fitness functions to guide an evolutionary run and this way evolve a series of stencils. After selecting the stencils from the framework archive, they were tested in the generator of portraits in order to assess to what extent (i) the input image remains recognisable in the typographic portrait and (ii) the input text remains legible. Figure 13 shows one of the stencils used in the creation of the portraits.

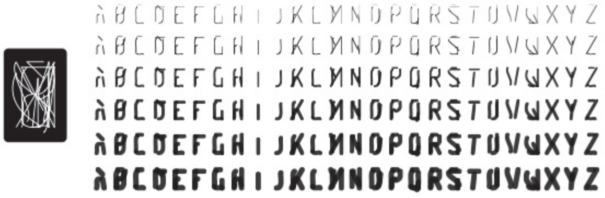


Figure 13: Stencil evolved and applied in experiment IV. The variation of the thickness of the stencil lines (left) generates a wide, continuous range of typographic weights that provide different visual emphasis (right).

A detailed description of the computational system that generates the typographic portraits is beyond the scope of this paper. Nevertheless, it is worthwhile summarising the main steps for the generation of each typographic portrait: (i) the input image is converted to grayscale; (ii) the brightness value of each pixel is calculated; (iii) an input text is composed from left to right and from top to bottom within a rectangular area proportional to the input image; (iv) for each glyph, the average brightness of the pixels located inside its bounds is calculated; and (v) the typographic weight of each glyph is set inversely proportional to the average brightness just calculated, *i.e.*, a glyph positioned over a dark area of the input image will be thicker than a glyph positioned over a lighter area.

The system that generates the portraits can be configured at different levels, *e.g.* number of text lines, leading, space between glyphs, width of the glyphs, minimum and maximum thickness of the glyphs. The adjusting of these parameters enables the generation of typographic portraits with different visual characteristics. For example, increasing the number of text lines provides more detail to the portrait; decreasing the leading and/or space between the glyphs makes the portrait visually denser; increasing the difference between the minimum and maximum thickness of the glyphs provides more contrast to the portrait. Furthermore, the system exports the generated portraits to vectors graphics. This way, one can print a postcard or a poster of his/her own typographic portrait.

Figure 14 shows typical typographic portraits created in this experiment. One can visualise animated versions of typographic portraits at http://cdv.dei.uc.pt/2018/complexity/portrait.mov. The obtained outcomes demonstrate that the stencils are able to maintain legibility and visual coherence among their glyphs while their thickness varies, which is important from a type design point of view.

The automatic evolution of new stencils enables the generation of unique typographic portraits. This reveals the potential of the evolved stencils for open-ended design projects, enabling the on-demand generation of unique typefaces.

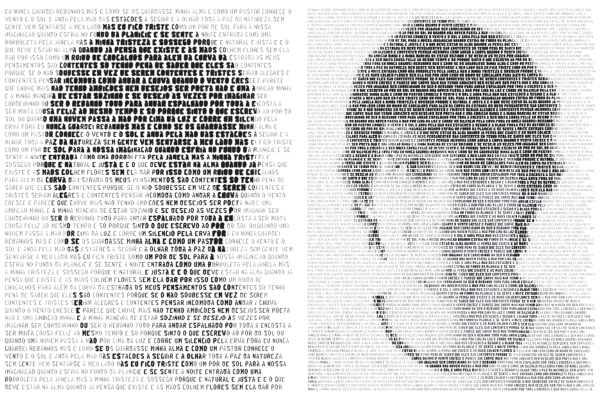


Figure 14: Typographic portraits composed of glyphs drawn with a stencil evolved with the presented framework. Two possibilities for the number of text lines are shown: 50 lines (left) and 100 lines (right). The input image is a photo of Sérgio Rebelo (graphic designer and researcher at CISUC), who tested the framework in experiment IV; and the input text is the poem "Eu Nunca Guardei Rebanhos", written in 1914 by Alberto Caeiro, an alter ego of Fernando Pessoa.

This experiment, which has been presented in a typography conference [38], demonstrates the application of the outcomes of the presented framework in a real design project, namely in the generation of typographic portraits.

4. Conclusion

An evolutionary framework for the automatic generation of type stencils was presented. We conducted a series of experiments to explore and assess this framework. Overall, the experimental results show the adequacy of the proposed framework to evolve stencils that (i) produce legible and coherent glyphs and (ii) are consistent with the preferences expressed by the user using the fitness function design interface. The results indicate that the approach guided by automatic fitness functions based on ML tend to optimise the fitness function. The results also revealed that optimising the different objectives of the fitness function will lead to legible and recognisable fonts but not necessarily aesthetically appealing. This behaviour is due to differences between how the ML techniques employed and humans perceive the inputs. Therefore, although legibility and recognisability are fundamental criteria for evolving glyphs, other visual aspects should be considered during evaluation in order to improve their aesthetic appeal.

This work demonstrated how EC and ML can inform contemporary design practices. The result is a framework that intends to provide alternative designs as stimuli for inspiration, working in a mind-opening way and promoting new ideas to create custom typefaces and letterings. This is useful, especially, when there will always be designers willing to experiment with the creation of fonts and to pursue new forms of typographic expression.

Future work will focus on: (i) implementation of adaptive mechanisms to the novelty search and archive mechanisms; (ii) creation of an archive interface to enable the user to change the behaviour of the archive mechanism during the evolutionary runs; (iii) enabling the user to save and insert evolved stencils into the evolutionary process; and (iv) implementation of the framework as web application and this way enable anyone to experiment with the evolution of type stencils.

Data Availability

The experimental results data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Funding

The first author is funded by Fundação para a Ciência e Tecnologia (FCT), Portugal [SFRH/BD/105506/2014]. We would also like to express our gratitude to NVIDIA for providing us one Titan Xp GPU to support our research.

References

[1] E. Lupton, *Thinking with Type: A Critical Guide for Designers, Writers and Students*, Princeton Architectural Press, 1st edition, 2004.

[2] I. Butterfield and M. Lewis, *Evolving Fonts*, 2000.

[3] A. Lund, "Evolving the shape of things to come: a comparison of direct manipulation and interactive evolutionary design," in *Proceedings of the International Generative Art Conference, Milan, Domus Argenia Publisher*, Rome, Italy, 2000.

[4] G. Levin, J. Feinberg, and C. Curtis, *The Alphabet Synthesis Machine*, 2001.

[5] T. Unemi and M. Soda, "An IEC-based support system for font design," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, pp. 968–973, Wash, D.C., USA, 2003.

[6] M. Schmitz, *GenoTyp, An Experiment about Genetic Typography*, 2004.

[7] M. Kuzma, "Interactive evolution of fonts," Tech. Rep., Technical University of Košice, 2008.

[8] K. Yoshida, Y. Nakagawa, and M. Köppen, "Interactive genetic algorithm for font generation system," in *Proceedings of the World Automation Congress '10*, pp. 1–6, TSI Press, 2010.

[9] T. Martins, J. Correia, E. Costa, and P. Machado, "Evotype: evolutionary type design," in *Evolutionary and Biologically Inspired Music, Sound, Art and Design*, vol. 9027 of *Lecture Notes in Computer Science*, pp. 136–147, Springer International Publishing, Cham, Switzerland, 2015.

[10] T. Martins, J. Correia, E. Costa, and P. Machado, "Evotype: from shapes to glyphs," in *Proceedings of the Genetic and Evolutionary Computation Conference '16*, pp. 261–268, ACM, New York, NY, USA, 2016.

[11] T. Martins, J. Correia, E. Costa, and P. Machado, “Evotype: towards the evolution of type stencils,” in *Computational Intelligence in Music, Sound, Art and Design*, vol. 10783 of *Lecture Notes in Computer Science*, pp. 299–314, Springer International Publishing, Cham, Switzerland, 2018.

[12] P. Machado, T. Martins, H. Amaro, and P. H. Abreu, “An interface for fitness function design,” in *Proceedings of the 3rd International Conference on Evolutionary and Biologically Inspired Music and Art*, J. Romero, J. McDermott, and J. Correia, Eds., Springer, Granada , Spain, 2014.

[13] P. Machado, T. Martins, H. Amaro, and P. H. Abreu, “Beyond interactive evolution: expressing intentions through fitness functions,” *Leonardo*, vol. 49, no. 3, pp. 251–256, 2016.

[14] J. Mouret, “Novelty-based multiobjectivization,” in *New Horizons in Evolutionary Robotics*, vol. 341 of *Studies in Computational Intelligence*, pp. 139–154, Springer Berlin Heidelberg, Berlin, Germany, 2011.

[15] A. Liapis, G. N. Yannakakis, and J. Togelius, “Sentient Sketchbook: Computer-aided game level authoring,” *FDG*, pp. 213–220, 2013.

[16] A. Vinhas, F. Assunção, J. Correia, A. Ekárt, and P. Machado, “Evolutionary and biologically inspired music, sound, art and design,” in *Proceedings of the 5th International Conference, EvoMUSART ’16*, C. Johnson, V. Ciesielski, J. Correia, and P. Machado, Eds., Lecture Notes in Computer Science, pp. 225–240, Springer International Publishing, Cham, Switzerland, 2016.

[17] E. Kindel, “The ’Plaque Découpée Universelle’: a geometric sanserif in 1870s Paris,” in *Typogr. Pap. 7*, pp. 71–80, Hyphen Press, The Department of Typography & Graphic Communication, University of Reading, 2007.

[18] J. Craig, I. K. Scala, and W. Bevington, *Designing with Type: The Essential Guide to Typography*, Watson-Guptill Publications, 5th edition, 2006.

[19] A. E. Eiben and J. E. Smith, *Introduction to Evolutionary Computing*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

[20] P. Machado, A. Vinhas, J. Correia, and A. Ekárt, “Evolving ambiguous images,” in *Proceedings of the IJCAI International Joint Conferences on Artificial Intelligence*, 2015.

[21] J. Correia, T. Martins, P. Martins, and P. Machado, “X-Faces: the exploit is out there,” in *Proceedings of the Seventh International Conference on Computational Creativity (ICCC ’16)*, F. Pachet, A. Cardoso, V. Corruble, and F. Ghedini, Eds., pp. 164–182, Sony CSL, Paris, France, 2016.

[22] J. Romero, P. Machado, A. Santos, and A. Cardoso, “On the development of critics in evolutionary computation artists,” in *Proceedings of the Workshops on Applications of Evolutionary Computation*, Springer Verlag, Essex, UK, 2003.

[23] J. Correia, P. Machado, J. Romero, and A. Carballal, “Evolving figurative images using expression-based evolutionary art,” in *Proceedings of Fourth International Conference on Computational Creativity (ICCC ’13)*, pp. 24–31, Creat, 2013.

[24] P. Machado, J. Correia, and J. Romero, “Expression-based evolution of faces,” in *Proceedings of the International Conference on Evolutionary and Biologically Inspired Music and Art EvoMUSART ’12*, P. Machado, J. Romero, and A. Carballal, Eds., vol. 7247 of *Lecture Notes in Computer Science book series*, pp. 187–198, Springer Verlag, 2012.

[25] C. Szegedy, W. Liu, Y. Jia et al., “Going deeper with convolutions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR ’15*, pp. 1–9, USA, 2015.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Proceedings of the Conference on Neural Information Processing Systems ’12*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., pp. 1097–1105, Curran Associates, Inc., 2012.

[27] K. Fukushima, “Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position,” *Biological Cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[28] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2323, 1998.

[29] S. Baluja, D. Pomerleau, and T. Jochem, “Towards automated artificial evolution for computer-generated images,” *Connection Science*, vol. 6, no. 2-3, pp. 325–354, 1994.

[30] P. Machado, J. Correia, and J. Romero, “Improving face detection,” in *Proceedings of the 15th European Conference on Genetic Programming (EuroGP ’12)*, A. Moraglio, S. Silva, K. Krawiec, P. Machado, and C. Cotta, Eds., pp. 73–84, Springer Berlin Heidelberg, Málaga, Spain, 2012.

[31] T. Kohonen, *Self-Organizing Maps*, Springer, Berlin, Germany, 3rd edition, 2001.

[32] T. Kohonen, “Essentials of the self-organizing map,” *Neural Networks*, vol. 37, pp. 52–65, 2013.

[33] L. Wang, Y. Zhang, and J. Feng, “On the Euclidean distance of images,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1334–1339, 2005.

[34] A. A. Goshtasby, “Similarity and dissimilarity measures,” in *Image Registration*, Advances in Pattern Recognition, pp. 7–66, Springer, London, UK, 2012.

[35] J. Lehman and K. O. Stanley, “Exploiting open-endedness to solve problems through the search for novelty,” in *Proceedings of 11th International Conference on Artificial Life (ALIFE XI ’08)*, MIT Press, Cambridge, Mass, USA, 2008.

[36] B. Tullett, *Typewriter Art: A Modern Anthology*, London, UK, Laurence King, 2014.

[37] F. Dietrich, “Visual intelligence: the first decade of computer art (1965-1975),” *Leonardo*, vol. 19, pp. 159–169, 1986.

[38] S. Rebelo, T. Martins, J. Bicker, and P. Machado, “Typography as image: experiments on typographic portraits,” in *Proceedings of the 9th Typography Meeting*, Instituto Politécnico de Tomar, Tomar, Portugal, 2018.

Hindawi
Complexity
Volume 2019, Article ID 3495962, 14 pages
https://doi.org/10.1155/2019/3495962

Research Article

Understanding Aesthetics and Fitness Measures in Evolutionary Art Systems

Colin G. Johnson,[1] **Jon McCormack**,[2] **Iria Santos**,[3] **and Juan Romero**[3]

[1] *School of Computing, University of Kent, Canterbury, UK*
[2] *SensiLab, Monash University, Caulfield East, VIC 3145, Australia*
[3] *University of A Coruña, A Coruña, Spain*

Correspondence should be addressed to Colin G. Johnson; c.g.johnson@kent.ac.uk

Received 23 November 2018; Revised 16 February 2019; Accepted 28 February 2019; Published 20 March 2019

Academic Editor: Guang Li

One of the general aims of evolutionary art research is to build a computer system capable of creating interesting, beautiful, or creative results, including images, videos, animations, text, and performances. In this context, it is crucial to understand how fitness is conceived and implemented to explore the "interestingness," beauty, or creativity that the system is capable of. In this paper, we survey the recent research on fitness for evolutionary art related to aesthetics. We also cover research in the psychology of aesthetics, including relation between complexity and aesthetics, measures of complexity, and complexity predictors. We try to establish connections between human perception and understanding of aesthetics with current evolutionary techniques.

1. Introduction

An ancient dream of humanity is to create models of itself. Ada Lovelace, often attributed as the first computer programmer, proposed to use computers for artistic tasks. Such tasks constitute a "grand challenge" since they present a series of subjective, social, and emotional characteristics that are often considered exclusive to human cultures.

One of the main difficulties in addressing this challenge is in developing formal models of human aesthetic preference. Such models would allow computer systems to predict the aesthetic taste of a human being or adapt to the aesthetic tendencies of a human group: in simple terms, to be able to make aesthetic evaluations and choices.

The term "aesthetic" derives from the Greek *aisthesis*, denoting feeling or perception, and its original meaning referred to sensory impressions. In the 18th century, it acquired a new meaning when Baumgarten's "Meditationes Philosophicae de Nonnullis ad Poema Pertinentibus" was published in Germany. This identified the relation between sensory experience and knowledge and gave the study of the knowledge of beauty the name *aesthetics*. From this moment, the term aesthetics is not restricted to the arts, but many of the things and experiences encountered in daily life. Hence, aesthetic decisions affect many many aspects of human choice and action, beyond those traditionally associated with fine art, for example.

Computational aesthetics (CA) can be defined as "the research of computational methods that can make applicable aesthetic decisions in a similar fashion as humans can." [1]. There are several papers that survey approaches to computational aesthetics, e.g., [2–4]. The term "computational aesthetics" is sometimes used in the sense of describing a particular class of artefacts made by computers, e.g., computer design and generative systems. However, in this paper, we will refer to computational aesthetics only as computational models of human aesthetics.

CA and the psychology of aesthetics (PA) have studied human aesthetics using a variety of different approaches. In this paper, we attempt to establish connections between these different approaches.

In the first section, we analyse several aesthetic modes included in recent evolutionary computation systems. The second section explores research results from psychology of aesthetics that will be of interest to AI researchers. Finally, we then propose some connections between the efforts of human psychology and AI and outline the advantages of this collaboration for both groups.

2. Aesthetic Fitness in Evolutionary Art

Since classical times, if not before, philosophers have engaged in the study of aesthetics: attempting to understand the nature of art and its appreciation, and why people engage in a specific set of *aesthetic behaviours* around artworks and make particular statements that can be called *aesthetic judgements*. In more recent times, these philosophical investigations have been joined by experimental and observational methods from psychology, neuroscience, and cognitive science, as well as research taking a constructive stance by building machines that produce work of aesthetic value, or machines that can themselves exhibit aesthetic behaviour and make aesthetic judgements.

A number of major strands exist in aesthetic theory [5, 6]. Early classical work focused on attempts to understand the nature of beauty—not just pointing out specific examples of beauty but identifying those aspects of objects that give rise to aesthetic appreciation. The earliest theories focused on the appreciation of skill in imitating the physical world, but later theorists found this lacking. In particular, once mechanical means for creating high-quality imitations were available such as photography and sound recording, new theories to explain the difference between simple reproductions and objects of aesthetic value became needed.

One major strand of aesthetic theory is based around the idea that aesthetic appreciation is a deliberate *act of expression* by the art-maker. In these theories, the art-maker is concerned with transmitting experiences and emotions that they have experienced to the audience for their work, in a way that cannot be readily done using more direct means such as purely descriptive text or diagrams. This gives rise to an immediate problem for computer art systems, which have no emotional qualia to form the grounds for expression. Nonetheless, even for machine-made art, we can sometimes recover some value from such expression-focused theories. One kind of computer art that can be said to be expressive is that which exploits the fact that computers are now almost universally networked systems, and for the expression to be an expression of the zeitgeist around a particular area as discovered online. For example, one version of the *Painting Fool* system [7] uses newspaper articles as the material that it "expresses" in a visual art form; the importance and salience of the source material come from the fact that is important enough to form a newspaper headline. Another way in which we can see expression explaining the impact of computer art systems is in those systems that act as a shaper and reinforcer of the user's interactions: not acting as an expressive device of the computer's own (absent) feelings, but allowing the user to explore and reinforce their expressions in a way that is not possible without the machine.

Another major strand of aesthetic theory is concerned with ideas of form. These theories argue that what makes an aesthetically engaging object distinct from a mundane one are formal aspects such as the placement of objects in an image, the use of symmetry, and the balance between order and complexity. The content is less relevant—aesthetic objects still have content, of course, but broadly similar content arranged without regard to form will have little aesthetic interest. Such theories are appealing to explain how computer art systems can create aesthetically engaging objects, because aspects of form can be encoded algorithmically, measures of form can be used as fitness drivers within learning and evolution systems, and different aspects of form can be brought together using multicriteria optimisation.

Another strand argues for the importance of social interactions and the social construction of aesthetic value, whether within a specific social discourse around art (e.g., Danto's [8] idea of the *Artworld*) or by being influenced by, and influencing, wider social and political issues. This has been occasionally explored in computer art systems [9, 10], by modelling a wider network of systems that create art and systems that critique and contextualise that art; however, there is much opportunity for more work in this area.

More recently, the focus has shifted from the wider world to the inner world of the brain and nervous system, examining the brain during aesthetic experiences [11]. Again, there are opportunities for this to be used in the context of fitness drivers for evolutionary art systems by modelling the potential audience response to art, much as user modelling [12] models the user response to more prosaic systems.

In contrast with theories that argue that aesthetics is a social phenomenon, other philosophers of aesthetics have taken a position that there are—at least at a very high level—some common features to aesthetic objects and to the act of aesthetic appreciation that remain constant over time. Dutton [13], for example, lists seven "aesthetic universals" that he claims form a feature of most social practices that are regarded as art. These are that

(i) the production of art objects requires skill and expertise

(ii) the objects give pleasure in-and-of themselves, regardless of whether they satisfy a practical need

(iii) art is produced in styles that are socially developed and are primarily about form and composition

(iv) art exists in the context of a critical and analytical discourse

(v) art objects imitate or symbolise aspects of the wider world

(vi) art objects are the subject of a special kind of attention and evoke particular behaviours towards them

(vii) audiences engage in art by using their faculties of imagination, and that artists make use of imagination in creating and developing artistic ideas and objects.

There is not necessarily a conflict between the idea of universals and the idea of social construction of aesthetics. It could be argued that whilst the broad categories of concepts that characterise art and aesthetic behaviour are broadly universal, specifics vary with time in a socially constructed way. Indeed, a major model of aesthetic appreciation and aesthetic judgement developed by Leder and colleagues uses an information-processing relationship between components that integrate into an aesthetic episode [14, 15]. The model includes low-level "universal" aesthetic properties, such as symmetry, complexity, contrast, and grouping, but also social, cognitive, and emotional components that all contribute in forming an aesthetic judgement.

2.1. Evolutionary Art Systems. Evolutionary art systems are computer systems that employ evolutionary computation (EC) methods to generate artworks [16]. Evolutionary art systems have been devised to create drawings, designs, buildings, poetry, sounds, music, 3D forms, images, and even choreography. Typically, the way in which these systems vary from other applications of EC is in the fitness function; other aspects of EC (selection methods, crossover and mutation operators, etc.) are largely the same as in more traditional optimisation applications. Such fitness functions can give rise to aesthetic value in two main ways. The first is *explicitly*, where the fitness function drives the evolutionary search towards items of greater aesthetic value. The second is *endogenously*, where the fitness creates a process that is itself of aesthetic value. An example of the latter is the body of work in artificial life art and artwork based on simulated ecology [17]. These might reflect a shift back towards an aesthetics of imitation in a new way—by simulating processes that occur on a temporal or spatial scale that is inaccessible to naked-eye viewing, they imitate/represent natural processes in a scaled or abstracted way making them accessible to immediate perceptual apprehension. This allows unspecialised audiences to reflect on these processes which are otherwise only comprehensible to scientists.

An evolutionary system that aims to generate aesthetically engaging material explicitly should therefore have a fitness function that drives the evolution toward areas of a search space that are aesthetically valued. So, the fitness function should be grounded in some theory of aesthetics; perhaps one of the established theories, or perhaps a new kind of theory that is distinctive to computer art or evolutionary art. Johnson [18] reviews a number of possible ideas on which such fitness functions could be built.

The most direct way to do this is via some kind of *aesthetic measure*. That is, the fitness function directly enacts some algorithmic method of scoring or ranking the aesthetic value of a specific work. This fits particularly well with aesthetic theories based around form—the most typical measures used are measures of formal aspects such as symmetry and complexity. In our discussion below on the psychology of aesthetics, we will see that this is the dominant theory there too; much of the experimental work in this area explores correlations between formal aspects of visual images and the viewer's aesthetic or affective responses.

One of the most influential EC-art papers in recent times that uses aesthetic measure is that of den Heijer and Eiben [19], which compares four different aesthetic measures as fitness functions for a EC system. The paper shows the results of the different functions in using them as fitness measures with an EC and by calculating the cross-evaluation of each function with the others. However, the problem with this type of approach is that whilst the functions proposed are useful as tools to explore the capabilities of EC, their connection to human aesthetic judgement is not clearly explained prior to their being employed as fitness functions. In some cases, the functions (called "measures" in the paper) were employed as metrics in learning systems, so they can be used for aesthetic purposes, but not necessary alone. In fact, one of the metrics analysed in the paper—the one first proposed by Machado and Cardoso [20]—was designed for monochrome images but applied in this research to colour ones.

Another way to create this fitness function is via a corpus of examples, typically in the form of an *inspiring set* [21] of examples that the computer system should use to inspire work that is new but in a similar style. The features provided to the learning system from the examples will dictate what aesthetic theories are underpinning this use of the corpus. For example, a system that uses geometrical analysis of the corpus examples, or extracts features based on the histogram of colours in the image, is driving towards aesthetics based around form or colour distribution. By contrast, if the system were using sentiment analysis to extract emotional cues from the corpus, this can be seen as working closer to expression and perceived emotion theories.

Another way to assign fitness is for the system to use interaction with people in place of a fixed function [22]. In terms of aesthetic theories, this leaves the theory to the user—rather than a computational fitness function being used, the decision on fitness is referred to a human, who can apply their own aesthetic judgement without having necessarily to theorise it formally. One under-explored area for future work would be for the human making the judgement to provide a more detailed critique of the work rather than just a selection or score, in some computer-readable form. This fits into a recent trend in evolutionary computation which uses richer *fitness drivers* containing much more information than a simple score or ranking [23] for selection and focused mutations. We can see this as fitting into a more social, critic-based theory of aesthetics, where human critics engage in a discourse with established or emerging artwork traditions.

3. Some Findings from the Psychology of Aesthetics

There is some overlap between current research on PA and CA. As an example, there are some researchers in PA looking for measures of aesthetic value or visual complexity. But at the same time, looking at the cross-citation of both areas, there is little communication between them. This section will analyse some of the findings in PA from the point of view of an AI researcher. We hope that this can help in creating computer systems that work with concepts such us visual complexity, aesthetics, and symmetry.

Firstly, a set of PA experiments only done with human beings are explored that relate aesthetic judgements to the complexity of the work produced. Next, we explore briefly some works that employ algorithmic measures of complexity, and other works that try to model visual complexity. Then, we review research that relates measurable properties of images to visual perception in the form of fractal analysis. Finally, we

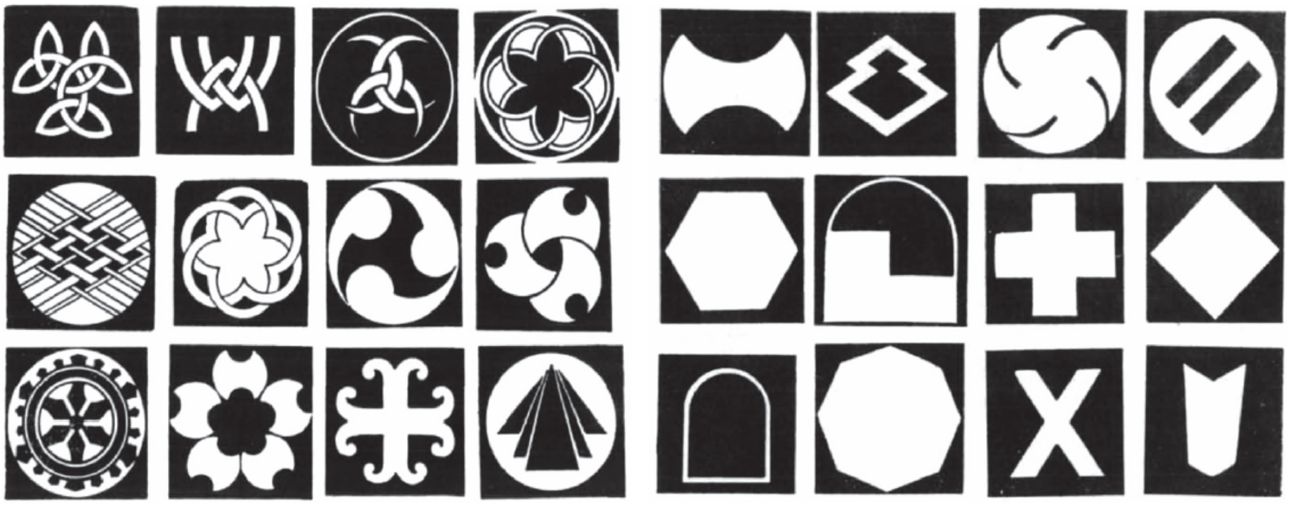


Figure 1: The twelve most-liked (left) and least-liked (right) of Eysenck's [33] experiments.

focus on possible aesthetic tests from PA that could be useful in AI research.

Before researchers came up with practical experiments in the psychology of aesthetics, questions related to art and aesthetics were answered by means of theories and experiences of the theorists themselves. The majority of cases were based exclusively in the observation of the reactions of a few viewers contemplating artistic works. Such informal processes, whilst useful for clarifying ideas, do not provide a strong basis for implementable theories.

In 1876, Fechner published "Elements of Aesthetics" [24], where he described a study based on the observation of the diverse answers of representative subjects of distinct populations with different visual material. These investigations laid the foundations and experimental methods for hypothesis formulation in aesthetics and its verification under controlled conditions.

Once this experimental basis had been established [25], the next step was to determine a method that was able to quantify the aesthetics of an object. The complex dimensions of a work of art, for example (form and location of the lines, rhythmic sequences, variations of tone, etc.), are from this moment objects of measurement: first the mathematician Birkhoff and later Eysenck would propose the first formulas for aesthetic measure. They were used as a measure of the aesthetic "value" in a number of different experiments and, as we shall see below, with contradictory results.

3.1. Experiments on Visual Complexity and Aesthetic. In the 1930s, Birkhoff set out the first mathematical formula that was designed to measure aesthetic value. This formula asserts that, for visual objects, the aesthetic measure of the object (M) is related to its order (O) and complexity (C), specified in the following relationship:

$$M = \frac{O}{C}. \tag{1}$$

Equation (1) proposes that the aesthetic measure of an image is correlated with the order and simplicity/complexity of its visual stimuli. Together with the presentation of that formula, Birkhoff [26] defined complexity as an expression of multiplicity, such as the number of elements that make up an image, while the order describes the regularity of those elements (repetition and redundancy).

While Birkhoff provided many different visual examples, he did not carry out experiments to validate his hypothesis. Even so, there are several research papers on his theory, some of which offer widely differing results. On the one hand, Brighouse [27] and Meier [28] conclude that the theory of Birkhoff is empirically founded, while, on the other hand, Weber [29], Beebe-Center and Pratt [30], Davis [31], and Eysenck [32] are not in agreement with this hypothesis. The most complete study related to the Birkhoff hypothesis was carried out by Eysenck [32–34]. Previously, Eysenck himself had carried out experiments related to this theory, exposing his disagreement with it. In order to be able to provide an alternative measure, he performed his own experiment in a controlled environment. A total of 11,000 participants, including those with art studies and without them (artists, students, teachers, and psychologists), were shown different series of polygons and asked to sort them according to their aesthetic preferences. These polygons were part of the material provided by Birkhoff [26]. From the experiment, Eysenck presents a formula different from that of Birkhoff, although also based on ideas of order and complexity. In this case, the relationship with complexity is positive, since both order and complexity were found to positively contribute to the appreciation of beauty.

It should be noted that the images employed by both Birkhoff and Eysenck are images with set of polygons, created for the experiment (not real-world images) and that both researchers do not have exact, much less computational measures that allow quantifying order or complexity. In the Figure 1 it can be seen some images employed in Eysenck experiments.

Berlyne [35] proposed that judgements about the interest and liking of an image depend, fundamentally, on the judgement of the complexity of that stimulus [36]. This, in turn, is related to factors such as the regularity of the model, the number of elements that make up the scene, its heterogeneity, or the irregularity of the forms [37]. The optimum of aesthetic pleasure would be latent until a subject encountered stimuli of average complexity, in case of having a very moderate stimulation potential, or stimuli that imply a very high potential, but reducible by appropriate modifications. This optimum varies according to learning [38].

Aesthetic preferences and judgements of beauty have been the subject of numerous research experiments since the formulation of order and complexity by Berlyne. Their hypotheses have been the subject of study following two different approaches: one based on general visual stimuli and another on artistic stimuli. In the case of visual stimuli, Aitken [39], Katz [40], and Vitz [41] use geometric objects while Heath et al. [42], Ichikawa [43], and Stamps III [44] perform their experiments with artificially generated images. With a focus based on artistic stimuli, we highlight the work carried out by Krupinski and Locher [45], Nicki and Moss [46], and Osborne and Farley [47] by means of abstract paintings, Nicki et al. [48] with works of Cubist art, Messinger [49] using figurative images, and Saklofske [50] by means of portraits.

The conclusions obtained across the experiments are contradictory, even within the same approach. Some find a distribution of preference in the form of an inverted U, with preference given to intermediate levels of complexity, whilst others observe a linear increase of aesthetic engagement with increasing complexity. A more detailed breakdown and analysis can be found in the paper by Nadal [51].

Berlyne himself [37] expressed a problem in conceptualisations of visual complexity. Attneave [52] and Berlyne [53] surveyed the subjective aspect of visual complexity. However, some experiments that use classification scales and other techniques confirm that collative variables and subjective information variables tend, as expected, to vary concomitantly with the corresponding objective measure of the classical theory of information [54]. Hogeboom explained that the complexity perceived by each individual depends on the way the scene is organized [55, 56]. This may be one of the reasons that the previous conclusions were contradictory.

Forsythe et al. [57] demonstrated that the subjective image complexity measure can be conditioned by familiarity. In Nadal et al. [58], a group of individuals rated the beauty and complexity of a set of images. The authors could not find any correlation between ratings. The researchers proposed three different types of complexity that can influence visual perception of complexity (asymmetry, the amount and variety of objects, and the way the objects are organised).

Also using ideas of priming and conditioning, Mallon et al. [59] studied the changes in the evaluation of the perceived beauty in abstract artworks and maintained that the perceived beauty increases after the exhibition of paintings that have been described as less beautiful and diminishes after the exhibition of paintings that were described as the most beautiful, which again reinforces the idea of subjectivity in aesthetic appreciation.

Güçlütürk et al. [60] call for a focus on individual differences in aesthetic preferences, and the adoption of alternative methods of analysis that take into account these differences, along with a reevaluation of the established rules of aesthetic preferences in humans. The relationship between aesthetic taste and stimulus complexity is commonly defined as an inverted U-shaped curve; images that are too simple offer too little to appeal to the aesthetic sense, whereas excessively complex images present too many diverse stimuli to allow aesthetically engaging patterns to be identified. However, frequent individual differences between the preferences of the participants' complexity have been observed since the first studies on the subject. The usual use of methods of linear analysis that ignore these great individual differences in aesthetic preferences gives an impression of high level of coincidence between individuals. In their study, they gather the qualities of taste and perception of the complexity of 30 participants for a set of 144 digitally generated grayscale images. In addition, an objective measure of the complexity of each image is calculated. The authors claim that the results show that the U-shaped relationship between the taste and the complexity of the stimulus is produced as the combination of different individual functions of taste. Specifically, after automatically grouping the participants in relation to their taste qualifications, they determine that a group of sample participants assigned increasingly lower quality of taste for more and more complex stimuli, while a second group of participants had scores of taste increasingly higher for more and more complex stimuli. The two groups differ as to whether they prefer complex or simple patterns, but not in the way in which they perceive the complexity. The group of participants who prefer the simplest patterns were faster in their taste assessments compared to the group that preferred complex patterns. These differences in the assessment time were not found in the evaluation of complexity. A partial explanation of the results is provided by the theory of fluidity of Reber et al. [61], according to which experience in fluid processing has a positive effect on the stimulus, so a decrease in taste towards complex stimuli could be expected (and therefore processed with less fluidity) compared to simpler stimuli (and processed more fluidly). This would validate the results of a group, but not those of the other.

A recent framework by Graf and Landwehr [62] called PIA (pleasure-interest model of aesthetic liking) aims to provide a better explanation of the contradictory patterns of preference for aesthetic stimuli that are easy or difficult to process. According to the authors, an aesthetic object can be processed in two stages. In the first stage, an automatic processing is carried out, and then, if the viewer is sufficiently motivated to continue the processing of the stimulus, there is a controlled processing. Similar to the theory of fluidity, the PIA model predicted that purely automatic processing of the stimuli results in a decrease in taste as the complexity of them increases. The prognostic model states furthermore that the controlled processing could give rise to an inverted U curve, if the levels of complexity of the stimuli are sufficiently high to cause disgust and confusion.

3.2. Measurements of Image Complexity. After exploring several PA ideas that try to analyse the relation between visual complexity and aesthetics using an ad hoc determination of complexity, we will move on in this section to survey some works that employ algorithmic measures of complexity.

As stated previously, perception of image complexity is subjective. The first method to calculate the complexity of a set of images is to relate complexity with another objective factor of the image. As an example, complexity could be related to the number of objects in an image. So, a constructed image of two triangles has a complexity of 2, while a constructed image with 9 triangles has a complexity of 9. Similar approaches use the number of *different* objects, and other objective qualities of constructed images. The first works analysed in the previous section employ this method by constructing the images used in the dataset (typically with combinations of polygons and other simple forms).

A different approach in order to determine the complexity of a image is to ask a group of people to self-report the perceived complexity and calculate the average of the responses. This gives a complexity measure for images that were not created specifically for the experiment (such us paintings or real-world photographs) [63–65]. This method was employed on most of the papers presented in the previous section. While this method is not limited to any specific kind of image, it may present a significant time or resource cost if the image corpus is large.

A computer generated measure of complexity can be applied to images with relatively little cost so it can be used to feed computer systems that generate images or other novel images [66]. Moreover, it can allow us to determinate the factors (emotional, semantic, etc.) that affect the human perception of image complexity, through a proportionate objective measure. Hence, it can be used to analyse the differences between objective and subjective values in different types of images. We will see a clear example of that later in the work of Jakesch and Leder [67]. Moreover, as we will see, some PA researchers such as Forsythe et al. [68] suggest that the objective measure (based on calculated metrics) can be more useful to predict human aesthetic preference than the subjective one (based on human scores).

Hochberg and Brooks [66] created a semiautomated measure of image complexity, based on the combination of number of interior angles, different angles, and lines. García et al. [69] developed an algorithm to measure the image complexity of icons using the number of lines (horizontal, vertical, and diagonal), forms (open and closed), and letters in each icon. Mcdougall et al. [70] employ the same measure for the complexity of a set of forms and achieve a correlation with the judgement of humans of Rs=0.73 for abstract icons.

Forsythe et al. [71] created an automatic system to measure the complexity of icons based on edge information and structural variability. They found high correlation between their scores and those provided by Garcia et al. (Rs=0.66 for edge information and Rs=0.65 for structural variability), and also for the studies of McDougall et al. (Rs=0.64 for edge information and Rs=0.65 for structural variability). To our knowledge, this system is the first example published in psychology that employs a computational metric to measure complexity.

In AI research, Machado and Cardoso [20] propose visual complexity metrics based on the compression rate and error of JPEG and Fractal compression. This was based on ideas from Arnheim [72–74] and Moles [75]. They base their measure of image complexity on findings from information theory. Other authors propose similar theories [76–78], where the complexity is related to the unpredictability of the image (of the pixels in the image) [79]. As a highly unpredictable image is not easy to compress, they used the length of the compressed file and the degree of error as estimates for the predictability of the image. The following equation shows the formulation of the measure:

$$Visual\_Complexity\_Measure = \frac{RMS\_Error}{Compression\_Ratio}. \quad (2)$$

In PA, Donderi and colleagues [80, 81] were also inspired by algorithmic information theory. They used JPEG and ZIP compression as an approximation of the minimum code to describe an image, as a consequence estimating the predictability of the image. An image with all pixels black is (in principle) easy to compress and is readily predictable. On the other hand, a random generated image with no relation between each pixel is not predictable at all and is also not compressible. In Donderi and McFadden [82], the authors get a correlation of Rs=0.77 between the length of JPEG and ZIP compressed files and subjective image complexity.

Forsythe et al. [57] presented four metrics based on perimeter, Canny, JPEG, and GIF compression. They tested these metrics with a number of previous datasets, showing high correlations with subjective complexity. In 2011, Forsythe et al. [83] analysed the correlation between the perceptual image complexity using several algorithmic measures: (i) length of JPEG compression, (ii) length of GIF compression, and (iii) perimeter detection measures. The authors employ a dataset of 800 images with 5 different categories: Abstract Artistic, Abstract Nonartistic, Representative Artistic, Representative Nonartistic, and Photographs. The results show a correlation of Rs=0.74 with the length of the GIF file for the Figurative Decorative category. Other categories had lower correlations (Abstract Decorative: Rs=0.6, Natural Pictures: Rs=0.55, Figurative Artistic: Rs=0.47, and Abstract Artistic: Rs=0.42).

Chikhman et al. [84] test different measures of complexity. With a dataset of 15 Chinese hieroglyphs, they found that the best measure is the "product of squared spatial-frequency median and the image areas." For a set of 24 outline images of objects, they found that the best measure is the number of turns in the image. Their conclusion is that different complexity estimates are needed for different types of images.

Marin and Leder [85] also analyse the correlation between computer-generated measures and perceptual image complexity. They use a subset of the International Affective Picture Systems (IAPS) [86], which contains a collection of images labelled with degrees of affective states that are expressed through those pictures. The correlation between length of the TIFF (Rs=0,53) and JPEG (rs=0,52) was higher than the one achieved using perimeter detection (Rs=0,44). The highest correlation found in this experiment was the RMS contrast, with a correlation of Rs=0.59. In a second experiment, done with a set of paintings, the correlation achieved was lower.

The differences in findings between Forsythe et al. [83] and Marin and Leder [85] could be explained by the datasets employed. The dataset in Forsythe et al. [83] contained images with highly differing complexity in five different categories. On the other hand, the two datasets of Marin and Leder [85] present less variation in complexity: the IASP dataset contains nonprofessional photographs designed for exploring different emotions and the dataset of paintings offers a very similar degree of complexity.

Using the same datasets as Marin and Leder [85], Marin et al. [87] analyse the effect of presentation time on perceptual complexity of images. Seventy women classified 96 images from IAPS dataset, presented each for 1, 5, and 25 seconds. The correlations between the objective measures and the subjective ones get higher with the longer exposition time. As before, the experiment with paintings was less conclusive.

Cavalcante et al. [88] propose the use of a combination of statistics of local contrast and spatial frequency as a measure of complexity. Their dataset contains 74 streetscape images from four cities, 40 daytime and 34 nighttime scenes. They compare the results of this metric with some of the state-of-the-art ones, including perimeter and JPEG complexity, finding that their proposed metric is the more robust regarding different time scenarios.

Jakesch and Leder [67] tested the role of ambiguity in human complexity perception. To do this, they employed artworks with high degree of ambiguity, and modifications of artworks with a low level of ambiguity. While both sets present similar results regarding computer measures (Jpeg, GIF, and perimeter detection), the perceptual complexity was different between the two sets. Humans considered those images with higher ambiguity to be more complex than the low ambiguity images.

Ciocca et al. [89] analyse the role of colour in complexity. They found that subjective scores for colour images present a high correlation to those of greyscale images, suggesting that colour is not related to perception of complexity. They use a range of image features but do not find any one capable of predicting image complexity.

Marin et al. [87] analyse the differences between three alternate ways to asses the 'hedonic tone' of an image: beauty, pleasantness, or liking. They used two datasets, one with 96 representational paintings and the other with 96 attractive environmental scenes converted into cartoons. The correlation between the three hedonic tone measures was higher in cartoons (Rs=0.85) than on paintings (Rs=0.73). With the dataset of paintings, correlation of complexity and beauty was Rs=0.26, with a "pleasantness" of Rs=-0.16 and not present for liking. In the cartoons dataset, correlation between complexity and the three hedonic tone measures was not found.

Friedenberg and Liby [90] analysed the correlation between beauty and compression metrics. The datasets contain images that are patterns of different density created for the experiment. They reported high correlations between beauty and GIF complexity (0.56) and contour length (0.47). They found no correlations between beauty and numbers of parts. Building on this work and using the same datasets, Gauvrit et al. [91] analysed the correlation between subjective beauty and several different complexity measures: density, number of blocks, GIF compression rate, edge length, entropy, and algorithmic complexity. They found that the participants tend to have a preference for some types of complexity, but not for all. That can explain partially the differences between reported results related to image complexity. The authors propose that researchers should specify which notion of complexity is behind each work.

Forsythe et al. [68] evaluated human scores for beauty, complexity, familiarity, and encounter. The authors calculated two automatic measures of complexity based on GIF and JPEG compression. The results show a high correlation between automatic measures and human perception of complexity (Rs=.78 for GIF compression). The better predictor for human beauty was GIF complexity. The authors state that "The data reported here suggests GIF complexity contributed in a small way to perceptions of beauty, but that beauty has no significant relationship with human judgements of visual complexity or familiarity with an image". The authors consider computer measures more reliable and valid than human collected perceptions of complexity.

Following this line of research, Madan et al. [92] found that emotional arousal and valence influence image complexity ratings. They found a correlation between arousal and visual complexity of Rs=.50, which was attenuated with bias-aware instructions to Rs=.40. Also, Forsythe et al. [57] found that familiarity and learning also influence image complexity ratings.

3.3. Visual Complexity Prediction. In this section, we analyse several works that employ a set of metrics and a machine learning system to predict the visual complexity of images. Most of the systems are created by AI researchers but some are created by PA and CA researchers together, with one published in a psychology journal.

Machado et al. [93] is the first attempt to create an automatic predictor of image complexity based on a combination of metrics. The dataset employed is the one used in Forsythe et al. [83], consisting of 800 images in 5 different categories. In the first experiment, the individual correlation is calculated between a large set of computer generated measures and the average perceptual image complexity. Higher correlation was obtained using a canny edge filter, with Rs=0.77. JPEG compression achieved a correlation of Rs=0.74. In the second experiment, the large set of measures was fed into a machine learning system based on Artificial Neural Networks (ANNs), which form a predictor of complexity. The correlation between the best predictor and the subjective image complexity was Rs=0.83. Edge density and JPEG compression error were the strongest predictors of human complexity rates. The predictor error was 0.09 (0.4 in a scale 1-5). The error was higher on "Representational Artistic" and "Photographs of Natural and Man-made Scenes" images, possibly due to more semantic meaning than Abstract (Artistic and Nonartistic) and Representational Nonartistic image categories.

Ciocca et al. [94] used genetic programming to build an image complexity predictor, using four measures: roughness, number of regions, chroma variance, and memorability. They reported a correlation of Rs=0.890 on the training set, 0.728 on the validation set, and 0.724 on the test set, outperforming the results of each of the measures individually.

Gartus and Leder [95] calculate a wide range of computational measures of complexity and combine them using a random forest (a standard machine learning technique) to predict image complexity. The images were a set of abstract patterns from the set used by Gartus and Leder [95], with different numbers of triangles on a white background. The dataset contains 152 asymmetric and 76 symmetric patters for five types of symmetry. They found several computer metrics to have positive correlation with complexity. One metric based on GIF compression had the highest correlation with Rs=0.634 and mirror symmetry having a negative correlation of Rs=-0.578. Combining the metric based on GIF and mirror symmetry together, they reported a correlation of Rs=0.903.

3.4. Measuring Visual Concepts. In this section, we focus on different visual concepts that relate to visual aesthetics and how they can be modelled using metrics. We begin with fractal dimension, then on principles of symmetry, colour gradient, and low-level processing.

The first work we are aware of to relate fractal dimension and aesthetics is that of Aks and Sprott [96], who analysed the correlation between aesthetic preferences and (i) fractal dimension and (ii) Lyapunov exponent of abstract patterns. They found a preference for values of fractal dimension and Lyapunov exponents that are typical in natural objects.

Taylor et al. [97] analysed the fractal dimension of paintings by the artist Jackson Pollock. Later, Taylor et al. [98] demonstrated that the fractal dimension of Pollock's paintings increased almost linearly for a decade. From that moment, the fractal dimension was considered a measure related to the image complexity and was employed on both psychological studies of aesthetics and artificial intelligence applied to aesthetics.

Spehar et al. [99] found a consistent aesthetic preference for fractal images. They employed forced-choice method of paired comparison and used images with different fractal dimension. They use three different datasets: (i) natural images, (ii) simulated coastlines, and (iii) Pollock's images. The results showed a "consistent trend for aesthetic preference to peak within the fractal dimension range 1.3–1.5 for the three different origins of fractal image." The authors consider this range as typical for natural objects.

Taylor et al. [100] analysed different responses to fractal patterns (visual preferences to physiological responses) in the work of painter Jackson Pollock. Jones-Smith and Mathur [101], however, question the use of fractal dimension in the work of Pollock.

Street et al. [102] present a large scale analysis of aesthetic preferences involving fractal and complexity metrics. The dataset used was composed of 81 abstract monochrome fractal images. After calculating a series of complexity measures, they found a strong negative correlation between fractal dimension (FD) and GIF ratio complexity measure, $Rs = -0.93$. They also used two-alternative forced choice analysis (TAFC) and obtained demographic information (age, gender, and continent of residence) from each participant. The results suggest strong differences related to continent and gender: in these experiments, females consistently preferred complex images over males.

In Spehar et al. [103], the authors use a set of 27 synthetic fractal images: nine $1/f$ filtered greyscale images with spectral slopes ranging from 0.5 to 2.5 in increments of 0.25, their thresholded black and white images and edges only counterparts. In a second experiment, they employed two further variations of the filtered greyscale images, called 'mountain' (that simulate a binary view of a mountain) and 'terrain' (that simulates a satellite view of a field with altitude shown in greyscale). They found that the majority of participants exhibited a peak preference for the intermediate fractal-scaling characteristics while other participants exhibited either a linear increase (aprox 20%) in preference with increasing amplitude spectrum slope or a linear decrease in preference with increasing amplitude spectrum slope (aprox 20%). The different tendencies were highly stable across all image types.

In his Ph.D. Thesis, Patuano [104] applied fractal dimension to landscapes. In order to do that, he employ several preprocessing stages to create a binary version of the image (using edges, silhouette outline, etc.) and then applied the box-counting method. The measure with the highest correlation to human preference was the fractal dimension of the image's extracted edges.

In Viengkham and Spehar [105], a set of images of tree levels of fractal dimension (low, medium, and high) are presented to a group of people, who are asked to rate liking, pleasantness, complexity, and interestingness. The study includes three types of synthetic fractal images and seven types of paintings. In most of the categories, a majority of participants prefer images with intermediate fractal dimension, with 40.13% compared with 33.05% of low fractal dimension and 26.82% of high FD.

Zipf [106] proposed that many phenomena follow a distribution where the frequency of occurrence is inversely proportional to its rank in the frequency table. So, the largest city of a country has double the population of the second one, three times more than the third one, and so on. Zipf's distribution is usual in language, but it can also describe city population sizes in a country, the number of people watching TV channels, and so on. Manaris and colleagues employ this distribution as metrics for music in several works, e.g., [107, 108]. Machado et al. [93] obtain a correlation with visual perceptual complexity of Rs=0.64.

The histogram of oriented gradients (HOG) counts occurrences of gradient orientation in localised portions of an image. The Pyramid Histogram of Orientation Gradients (PHOG) contains the HOG of the image with HOGs of parts of the image. Redies et al. [109] propose two metrics based on PHOG: self-similarity and complexity. They calculated the metrics for different datasets and found that one of those datasets (containing images of art paintings) could be characterised by a specific combination of values of these metrics.

Lyssenko et al. [110] found a correlation between subjective visual complexity and (i) PHOG self-similarity ($Rs = 0.56$) and HOG complexity ($Rs = 0.682$). The dataset consisted of 79 abstracts artworks. They also found correlations between these metrics and subjective terms that participants use to describe the artworks.

There are some studies that have tried to establish relationships between aesthetic value and colour gamut. Nascimento et al. [111] analyse the effects of changing the colour gamut of paintings to increase their aesthetic value. They asked a group of users to change the colour gamut of ten paintings. The maximum of the distribution was the same as the original, suggesting, unsurprisingly, that the chromatic compositions of the paintings employed matched the viewers' preferences.

Other works have investigated the relation between aesthetics and symmetry. Weichselbaum et al. [112] tested symmetry preferences of participants over different levels of individual art expertise. They found that "with higher art expertise, the ratings for the beauty of asymmetrical patterns significantly increased, but, again, participants preferred symmetrical over asymmetrical patterns". Thömmes and Hübner [113] analysed the relation between Instagram "likes" and three computational measures: two measures of visual balance and the preference for curvature over angularity. They utilised 700 architectural photographs from Instagram accounts. They found a positive correlation between visual balance and likes in 3D photographs, and a negative correlation in 2D ones. To the best of our knowledge, it is the first work that employees "likes" as a measure of aesthetic appeal.

3.5. Psychological Testing Related to Aesthetics. There are a number of psychological tests related to aesthetic judgement. These tests are relatively objective and easy to reproduce and provide quantified results [51, 52, 55, 56, 114–117]. The main problem with these tests, however, is the lack of consensus about them. The validity of concepts behind each test are debatable, typically being based on aesthetic principles proposed by the author of the test, but not accepted universally. The results of individual tests also vary between different studies, maybe due to selection of participants and other exogenous factors. As an example, Weichselbaum et al. [112] show that artistic experience affects symmetry preferences.

Graves [118] developed the *Design Judgement Test*. This test is based on theories of artistic creation and appreciation [119]. The author claims that this test can estimate certain capabilities related to artistic and aesthetic evaluation. To do this, the test estimates the degree of reaction to specific principles of aesthetics (according to the author) such us unity, drive, predominance, variety, balance, continuity, symmetry, proportion, and rhythm. Such principals may not be universally accepted or applicable [120–122]. A test consists of ninety pages. Each page contains two or three similar designs. One of the designs obeys all the commented principles while the remaining ones break at least one of them. The task of the individual doing the test is to select those designs that do not break any of the principles.

The average results obtained by participants in this test vary between studies [121, 122]. Although this can be, at least partially, explained by the selection of participants and other exogenous factors, it makes it hard to understand what constitutes a good score in this test. In the test done by Graves, art students get a higher average score than students who did not study art [118]. Graves concludes that the test can be used to differentiate between those two groups. Eysenck and Castle [121] obtain very different results, showing only minor differences between artistic and no-artistic students (64,4% vs. 60%), and differences between males and females. Eysench explains that the different results regarding art students can be related to changes in artistic education that in 1971 promote more regularity and simplicity than in 1948. Götz and Götz [123] report that "22 different arts experts (designers, painters, sculptors) had 0.92 agreement on choice of preferred design, albeit being critical of them" [124].

Machado and Cardoso [20] propose an aesthetic measure based on processing complexity and image complexity: "images that are simultaneously visually complex and easy to process are the images that have higher aesthetic value." Fractals are an easy example of very complex images but easy to process due to the self-similarity. Using metrics based on compression described below, and a fixed equation for aesthetics, they obtain scores up to 66 (corresponding to a 73.3% success rate), which is larger than those obtained with fine art graduates. In Machado and Cardoso [125], the authors employ a similar equation as fitness for a genetic programming engine that creates images. Romero et al. [126] employ some metrics related to JPEG, fractal compression and Zipf's law, and an ANN-based machine learning system to predict the answer of the test, resulting in an accuracy of 74.49%, similar to the previous study.

Hayn-Leichsenring et al. [127] studied a relationship between objective image measures and the subjective evaluations of the JenAesthetics dataset. This dataset consists of 1628 high-quality images of paintings (http://www.inf-cv.uni-jena.de/en/jenaesthetics). The objective measures are low-level image statistics related to aesthetics in previous research, such as those from Braun et al. [128], related to selfishness, anisotropy, and complexity. The subjective evaluations were aesthetic (defined as artistic value) and beauty (defined as individual attachment). The results revealed that the paintings of each period present specific statistical properties of the images. Moreover, they show evidence of correlation between beauty and aesthetics, and correlation between aesthetics and some objective measures on different subsets. The highest correlation was found between self-similarity and the beauty of a subset of paintings of buildings with $Rs = 0.50$. They found differences between aesthetic and beauty scores.

4. Conclusions

The researches from PA and CA have several main differences. First, we analyse some differences regarding the datasets and the results. The datasets used in the PA experiments usually contain a small number of images, due to the need to evaluate each of them by a group of human participants. In CA, the ideal is to have a large dataset of images that can allow machine learning to have more complete and diverse information.

Some datasets used in CA work are based on website photographic collections that have a large number of contributed images [129–131]. However, the images in these datasets were evaluated online in an uncontrolled environment and may have potential biases depending on relations with the author of the image, popularity reinforcement, display environment, and so on. Finally, as the information (images and evaluations) was provided from photographic websites, it is not clear what the users are evaluating (photograph quality, originality, visual aesthetic, and liking). An interesting approach is to employ a game to obtain evaluations of images. That allows the researcher to provide clear choices for evaluation and may encourage participants to spend more time contributing to the research. In Hacker and von Ahn [132], the authors employ a two-player game where each participant should evaluate images following the taste of the other participant. They recruited thousands of players and have collected millions of judgements.

From PA research, we learn that, even in PA experiments done in controlled conditions, users provide substantially different evaluations depending on the term and context of the question [87, 127]. Hence, we propose that when using such website datasets it is necessary to experimentally test what users are evaluating. And if possible, the best way to proceed is to create datasets in collaboration with PA researchers, with evaluations done in controlled environments and with a number of images that support the use of machine learning techniques.

Many of the curated datasets of art images are restricted to Western or European art, raising issues of cultural bias. Likewise, many of the reported studies are undertaken in Europe or North America, which may impact the diversity of study participants. With increasing scrutiny on how AI datasets are obtained for machine learning applications, researchers need to be aware of implicit or explicit bias in their selection of training data. This is an ongoing issue for research in this field.

Finally, some recent PA research comments on the main differences between individuals in appreciation of visual aesthetics and complexity [60]. A more detailed analysis of this issue could be very interesting. Moreover, it can be interesting to create a large set of images with individual evaluations of human beings, allowing the training of computer system to evolve to the aesthetic preferences of one individual human.

Regarding the results, CA research typically reports results using a success rate or RMS error, while psychologists are more likely to use correlation. This is not a major problem: some papers get good results in correlation employing ML systems that try to minimise RMS error [93], but future systems trained to maximise correlation can achieve better results.

Closer collaboration between PA and CA can give rise to results that advance both disciplines. Given the general quality of datasets (PA), enormous sets of computer metrics (CA) and ML techniques (CA), and posterior analysis (both), more powerful predictors of visual complexity can be built. AI researchers can even use AI methods to build new metrics that no one was thinking about (using GP programming such as in Ciocca et al. [94], Artificial Neural Networks [133], or deep learning [134]). Predictors can be implemented that allow remote access (via a web page, for example), allowing any researcher to get a visual complexity value for an image or set of images online. This will help make the analysis of aesthetics and complexity objective metrics more accurate than individual feature analysis, and accessible for everyone. In this context, it is remarkable that the work of Forsythe et al. [68] finds more correlation between aesthetics with objective complexity measures (GIF compression metric) than with subjective complexity. It could be interesting to undertake a similar analysis with a complexity predictor made by a combination of metrics. Additionally, the detailed analysis of the predictor then allows us to know more information about the relevant metrics related to complexity. Obviously, the definition of some standards, one dataset, one complexity predictor, etc., that is acceptable for everyone will help in this schema of common research.

Some CA researchers begin the research with the idea of creating a computer system able to create original and aesthetically valuable artworks [135]. Generative techniques such us genetic programming are very interesting for this research because they can be used to illustrate the results of a metric or combination of them [19, 136]. However, better research results in complexity and aesthetic prediction are needed in order to advance image generation systems. Here, a collaboration between PA and CA is needed in order to achieve the new research results required. Moreover, even without being used in conjunction with generative systems, computer aesthetics systems can have enormous real-world applications.

Evolutionary art and computational aesthetics are relatively young areas of research. Yet, some authors may think that there is nothing new that can be done. Our purpose with this paper is to help the development of both areas by highlighting some possible underexplored pathways and to illustrate the exciting and valuable prior research from the psychology of aesthetics.

We hope for a future where several visual complexity and aesthetic predictors are accessible online, where evolutionary art tools are widely employed by people as ways of exploring their creative capacity, and where computer systems can convincingly create paintings in the style of any human artist and beyond.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors' Contributions

Colin G. Johnson, Jon McCormack, Iria Santos, and Juan Romero contributed equally to this work.

Acknowledgments

This work is supported by Xunta de Galicia (Ref. XUGA-PGIDIT-10TIC105008-PR), grant for the consolidation and structuring of competitive research units of the University System of Galicia: Competitive Reference Groups (ED431C 2018/49), grant from the Ministry of Education, Culture, and Sport for mobility stays of professors and researchers in foreign higher education centers and investigation (PRX18/00117), and an Australian Research Council Future Fellowship FF170100033.

References

[1] F. Hoenig, "Defining computational aesthetics," in *Proceedings of the First Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging, Computational Aesthetics '05*, pp. 13–18, Aire-la-Ville, Switzerland, Switzerland, 2005.

[2] G. Greenfield, "On the origins of the term "computational aesthetics"," in *Proceedings of the First Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging, Computational Aesthetics '05*, pp. 9–12, Aire-la-Ville, Switzerland, 2005.

[3] P. Galanter, "Computational aesthetic evaluation: steps towards machine creativity," in *Proceedings of the ACM SIGGRAPH 2012 Courses, SIGGRAPH '12*, pp. 14:1–14:162, New York, NY, USA, 2012.

[4] G. Birkin, *Aesthetic complexity: practice and perception in art & design [Ph.D. thesis]*, Nottingham Trent University, 2010.

[5] N. Carroll, *Philosophy of Art: A Contemporary Introduction*, Routledge, 1999.

[6] B. Gaut and D. M. Lopes, Eds., *The Routledge Companion to Aesthetics*, Routledge, 2013.

[7] A. Krzeczkowska, J. El-Hage, S. Colton, and S. Clark, "Automated collage generation—with intent," in *Proceedings of the International Conference on Computational Creativity*, D. Ventura et al., Ed., pp. 36–40, 2010.

[8] A. Danto, "The artworld," *The Journal of Philosophy*, vol. 64, no. 19, pp. 571–584, 1964.

[9] P. Machado, J. Romero, M. L. Santos, A. Cardoso, and B. Manaris, "Adaptive critics for evolutionary artists," in *Applications of Evolutionary Computing*, G. Raidl et al., Ed., vol. 3005 of *Lecture Notes in Computer Science*, pp. 437–446, Springer Berlin Heidelberg, 2004.

[10] J. Romero, P. Machado, A. Santos, and A. Cardoso, "On the development of critics in evolutionary computation artists," in *Applications of Evolutionary Computing: EvoWorkshops 2003*, S. Cagnoni, Ed., vol. 2611 of *Lecture Notes in Computer Science*, pp. 559–569, Springer, 2003.

[11] A. Chatterjee, *The Aesthetic Brain*, Oxford University Press, 2013.

[12] G. Fischer, "User modeling in human-computer interaction," *User Modeling and User-Adapted Interaction*, vol. 11, no. 1-2, pp. 65–86, 2001.

[13] D. Dutton, "Aesthetic universals," in *The Routledge Companion to Aesthetics*, B. Gaut and D. M. Lopes, Eds., pp. 267–278, 2013.

[14] H. Leder, B. Belke, A. Oeberst, and D. Augustin, "A model of aesthetic appreciation and aesthetic judgments," *British Journal of Psychology*, vol. 95, no. 4, pp. 489–508, 2004.

[15] H. Leder and M. Nadal, "Ten years of a model of aesthetic appreciation and aesthetic judgments: the aesthetic episode - developments and challenges in empirical aesthetics," *British Journal of Psychology*, vol. 105, no. 4, pp. 443–446, 2014.

[16] M. Lewis, "Evolutionary visual art and design," in *The Art of Artificial Evolution: A Handbook on Evolutionary Art and Music*, J. Romero and P. Machado, Eds., pp. 3–37, 2008.

[17] J. McCormack, "Creative ecosystems," J. McCormack and M. d'Inverno, Eds., pp. 39–60, Springer, Computers and Creativity, 2012.

[18] C. G. Johnson, "Fitness in evolutionary art and music: a taxonomy and future prospects," *International Journal of Arts and Technology*, vol. 9, no. 1, pp. 4–25, 2016.

[19] E. den Heijer and A. E. Eiben, "Comparing aesthetic measures for evolutionary art," in *Applications of Evolutionary Computation*, C. Di Chio, A. Brabazon, G. A. Di Caro et al., Eds., pp. 311–320, Springer Berlin Heidelberg, Berlin, Heidelberg, Germany, 2010.

[20] P. Machado and A. Cardoso, "Computing aesthetics," in *Proceedings of the 14th Brazilian Symposium on Artificial Intelligence: Advances in Artificial Intelligence, SBIA '98*, pp. 219–228, Springer-Verlag, London, UK, 1998.

[21] G. Ritchie, "Some empirical criteria for attributing creativity to a computer program," *Minds and Machines*, vol. 17, no. 1, pp. 67–99, 2007.

[22] H. Takagi, "Interactive evolutionary computation: fusion of the capabilities of EC optimization and human evaluation," *Proceedings of the IEEE*, vol. 89, no. 9, pp. 1275–1296, 2001.

[23] K. Krawiec, J. Swan, and U.-M. O'Reilly, "Behavioral program synthesis: insights and prospects," in *Genetic Programming Theory and Practice XIII*, R. Riolo, B. Worzel, M. Kotanchek, and A. Kordon, Eds., pp. 169–183, Springer, 2016.

[24] G. T. Fechner, *Vorschule der Aesthetik*, vol. 1, Breitkopf & Härtel, 1876.

[25] G. T. Fechner, "Zur experimentalen aesthetik," in *Abhandlungen der Mathematisch-Physischen Classe der Königlich Sächsischen Gesellschaft der Wissenschaften. IX. bd*, S. Hirzel, Ed., vol. 1, 1871.

[26] G. Birkhoff, *Aesthetic Measure*, Harvard University Press, 1933.

[27] G. Brighouse, "Variability in preferences for simple forms," *Psychological Monographs*, vol. 51, no. 5, pp. 68–74, 1939.

[28] N. C. Meier, *Art in Human Affairs; an Introduction to the Psychology of Art*, McGraw-Hill, 1942.

[29] C. O. Weber, "The aesthetics of rectangles and theories of affection," *Journal of Applied Psychology*, vol. 15, no. 3, pp. 310–318, 1931.

[30] J. G. Beebe-Center and C. C. Pratt, "A Test of Birkhoff'S Aesthetic Measure," *The Journal of General Psychology*, vol. 17, no. 2, pp. 339–353, 1937.

[31] R. C. Davis, "An evaluation and test of Birkhoff's aesthetic measure formula," *The Journal of General Psychology*, vol. 15, no. 2, pp. 231–240, 1936.

[32] H. J. Eysenck, *The Experimental Study of the Good Gestalt: A New Approach*, Lancaster Press, 1942.

[33] H. J. Eysenck, "'Type'-factors in aesthetic judgements," *British Journal of Psychology. General Section*, vol. 31, no. 3, pp. 262–270, 1941.

[34] H. J. Eysenck, "The empirical determination of an aesthetic formula," *Psychological Review*, vol. 48, no. 1, p. 83, 1941.

[35] D. E. Berlyne, "Complexity and incongruity variables as determinants of exploratory choice and evaluative ratings," *Canadian Journal of Psychology*, vol. 17, no. 3, pp. 274–290, 1963.

[36] D. E. Berlyne, J. C. Ogilvie, and L. C. Parham, “The dimensionality of visual complexity, interestingness, and pleasingness,” *Canadian Journal of Psychology*, vol. 22, no. 5, pp. 376–387, 1968.

[37] D. E. Berlyne, “Novelty, complexity, and hedonic value,” *Perception & Psychophysics*, vol. 8, no. 5, pp. 279–286, 1970.

[38] R. Frances, *Psicología del Arte y la Estética*, vol. 3, Ediciones AKAL, 1985.

[39] P. P. Aitken, “Judgments of pleasingness and interestingness as functions of visual complexity,” *Journal of Experimental Psychology*, vol. 103, no. 2, pp. 240–244, 1974.

[40] B. F. Katz, “What makes a polygon pleasing?” *Empirical Studies of the Arts*, vol. 20, no. 1, pp. 1–19, 2002.

[41] P. C. Vitz, “Preference for different amounts of visual complexity,” *Behavioural Science*, vol. 11, no. 2, pp. 105–114, 1966.

[42] T. Heath, S. G. Smith, and L. Bill, “Tall buildings and the urban skyline: the effect of visual complexity on preferences,” *Environment and Behavior*, vol. 32, no. 4, pp. 541–556, 2000.

[43] S. Ichikawa, “Quantitative and structural factors in the judgment of pattern complexity,” *Perception & Psychophysics*, vol. 38, no. 2, pp. 101–109, 1985.

[44] A. E. Stamps III, “Entropy, visual diversity, and preference,” *The Journal of General Psychology*, vol. 129, no. 3, pp. 300–320, 2002.

[45] E. Krupinski and P. Locher, “Skin conductance and aesthetic evaluative responses to nonrepresentational works of art varying in symmetry,” *Bulletin of the Psychonomic Society*, vol. 26, no. 4, pp. 355–358, 1988.

[46] R. M. Nicki and V. Moss, “Preference for non-representational art as a function of various measures of complexity,” *Canadian Journal of Psychology/Revue canadienne de psychologie*, vol. 29, no. 3, pp. 237–249, 1975.

[47] J. W. Osborne and F. H. Farley, “The relationship between aesthetic preference and visual complexity in absract art,” *Psychonomic Science*, vol. 19, no. 2, pp. 69-70, 1970.

[48] R. M. Nicki, P. L. Lee, and V. Moss, “Ambiguity, cubist works of art, and preference,” *Acta Psychologica*, vol. 49, no. 1, pp. 27–41, 1981.

[49] S. M. Messinger, “Pleasure and complexity: Berlyne revisited,” *The Journal of Psychology: Interdisciplinary and Applied*, vol. 132, no. 5, pp. 558–560, 1998.

[50] D. H. Saklofske, “Visual aesthetic complexity, attractiveness and diversive exploration,” *Perceptual and Motor Skills*, vol. 41, no. 3, pp. 813-814, 1975.

[51] M. Nadal, *Complexity and aesthetic preference for diverse visual stimuli [Ph.D. thesis]*, Departament de Psicologia, Universitat de les Illes Balears, 2007.

[52] F. Attneave, “Physical determinants of the judged complexity of shapes,” *Journal of Experimental Psychology*, vol. 53, no. 4, pp. 221–227, 1957.

[53] D. E. Berlyne, *Studies in the New Experimental Aesthetics: Steps toward an Objective Psychology of Aesthetic Appreciation*, Hemisphere, 1974.

[54] G. C. Cupchik and D. E. Berlyne, “The perception of collative properties in visual stimuli,” *Scandinavian Journal of Psychology*, vol. 20, no. 1, pp. 93–104, 1979.

[55] M. Hogeboom and C. Van Leeuwen, “Visual search strategy and perceptual organization covary with individual preference and structural complexity,” *Acta Psychologica*, vol. 95, no. 2, pp. 141–164, 1997.

[56] L. Strother and M. Kubovy, “Perceived complexity and the grouping effect in band patterns,” *Acta Psychologica*, vol. 114, no. 3, pp. 229–244, 2003.

[57] A. Forsythe, G. Mulhern, and M. Sawey, “Confounds in pictorial sets: The role of complexity and familiarity in basic-level picture processing,” *Behavior Research Methods*, vol. 40, no. 1, pp. 116–129, 2008.

[58] M. Nadal, E. Munar, G. Marty, and C. Cela-Conde, “Visual complexity and beauty appreciation: Explaining the divergence of results,” *Empirical Studies of the Arts*, vol. 28, no. 2, pp. 173–191, 2010.

[59] B. Mallon, C. Redies, and G. U. Hayn-Leichsenring, “Beauty in abstract paintings: perceptual contrast and statistical properties,” *Frontiers in Human Neuroscience*, vol. 8, p. 161, 2014.

[60] Y. Güçlütürk, R. H. A. H. Jacobs, and R. Van Lier, “Liking versus complexity: Decomposing the inverted U-curve,” *Frontiers in Human Neuroscience*, vol. 10, p. 112, 2016.

[61] R. Reber, N. Schwarz, and P. Winkielman, “Processing fluency and aesthetic pleasure: Is beauty in the perceiver’s processing experience?” *Personality and Social Psychology Review*, vol. 8, no. 4, pp. 364–382, 2004.

[62] L. K. M. Graf and J. R. Landwehr, “A dual-process perspective on fluency-based aesthetics: the pleasure-interest model of aesthetic liking,” *Personality and Social Psychology Review*, vol. 19, no. 4, pp. 395–410, 2015.

[63] P. Bonin, R. Peereman, N. Malardier, A. Méot, and M. Chalard, “A new set of 299 pictures for psycholinguistic studies: French norms for name agreement, image agreement, conceptual familiarity, visual complexity, image variability, age of acquisition, and naming latencies,” *Behavior Research Methods, Instruments, and Computers*, vol. 35, no. 1, pp. 158–167, 2003.

[64] F. X. Alario and L. Ferrand, “A set of 400 pictures standardized for French: norms for name agreement, image agreement, familiarity, visual complexity, image variability, and age of acquisition,” *Behavior Research Methods, Instruments, & Computers*, vol. 31, no. 3, pp. 531–552, 1999.

[65] Y. M. Cycowicz, D. Friedman, M. Rothstein, and J. G. Snodgrass, “Picture naming by young children: Norms for name agreement, familiarity, and visual complexity,” *Journal of Experimental Child Psychology*, vol. 65, no. 2, pp. 171–237, 1997.

[66] J. Hochberg and V. Brooks, “The Psychophysics of Form: Reversible-Perspective Drawings of Spatial Objects,” *The American Journal of Psychology*, vol. 73, no. 3, pp. 337–354, 1960.

[67] M. Jakesch and H. Leder, “The qualitative side of complexity: Testing effects of ambiguity on complexity judgments,” *Psychology of Aesthetics, Creativity, and the Arts*, vol. 9, no. 3, pp. 200–205, 2015.

[68] A. Forsythe, N. Street, and M. Helmy, “Revisiting Rossion and Pourtois with new ratings for automated complexity, familiarity, beauty, and encounter,” *Behavior Research Methods*, vol. 49, no. 4, pp. 1484–1493, 2017.

[69] M. García, A. N. Badre, and J. T. Stasko, “Development and validation of icons varying in their abstractness,” *Interacting with Computers*, vol. 6, no. 2, pp. 191–211, 1994.

[70] S. J. P. McDougall, M. B. Curry, and O. De Bruijn, “Measuring symbol and icon characteristics: Norms for concreteness, complexity, meaningfulness, familiarity, and semantic distance for 239 symbols,” *Behavior Research Methods, Instruments, and Computers*, vol. 31, no. 3, pp. 487–519, 1999.

[71] A. Forsythe, N. Sheehy, and M. Sawey, “Measuring icon complexity: an automated analysis,” *Behavior Research Methods, Instruments, & Computers*, vol. 35, no. 2, pp. 334–342, 2003.

[72] R. Arnheim, *Art and Visual Perception, a Psychology of the Creative Eye*, Faber and Faber, London, UK, 1956.

[73] R. Arnheim, *Towards a Psychology of Art/Entropy and Art — An Essay on Disorder and Order*, The Regents of the University of California, 1966.

[74] R. Arnheim, *Visual Thinking*, University of California Press, Berkeley, Califf, USA, 1969.

[75] A. Moles, *ThÉorie de L'Information et Perception EsthÉtique*, Denoel, 1958.

[76] E. L. Leeuwenberg, "Quantitative specification of information in sequential patterns," *Psychological Review*, vol. 76, no. 2, pp. 216–220, 1969.

[77] H. A. Simon, "Complexity and the representation of patterned sequences of symbols," *Psychological Review*, vol. 79, no. 5, pp. 369–382, 1972.

[78] J. Schmidhuber, "Facial beauty and fractal geometry," 1998, http://cogprints.org/690/.

[79] D. Salomon, *Data Compression: The Complete Reference*, Springer-Verlag, Berlin, Heidelberg, Germany, 2006.

[80] D. C. Donderi, PWGSC Contract No, and S. McFadden, *A Complexity Measure for Electronic Displays: Final Report on the Experiments*, Department of National Defence, Defence Research & Development Canada-Toronto, 2003.

[81] D. C. Donderi, "Visual complexity: a review," *Psychological Bulletin*, vol. 132, no. 1, pp. 73–97, 2006.

[82] D. C. Donderi and S. McFadden, "Compressed file length predicts search time and errors on visual displays," *Displays*, vol. 26, no. 2, pp. 71–78, 2005.

[83] A. Forsythe, M. Nadal, N. Sheehy, C. J. Cela-Conde, and M. Sawey, "Predicting beauty: fractal dimension and visual complexity in art," *British Journal of Psychology*, vol. 102, no. 1, pp. 49–70, 2011.

[84] V. Chikhman, V. Bondarko, M. Danilova, A. Goluzina, and Y. Shelepin, "Complexity of images: experimental and computational estimates compared," *Perception*, vol. 41, no. 6, pp. 631–647, 2012.

[85] M. M. Marin and H. Leder, "Examining complexity across domains: relating subjective and objective measures of affective environmental scenes, paintings and music," *PLoS ONE*, vol. 8, no. 8, Article ID e72412, 2013.

[86] P. J. Lang, "International affective picture system (iaps): affective ratings of pictures and instruction manual," Tech. Rep., 2005.

[87] M. M. Marin, A. Lampatz, M. Wandl, and H. Leder, "Berlyne revisited: evidence for the multifaceted nature of hedonic tone in the appreciation of paintings and music," *Frontiers in Human Neuroscience*, vol. 10, p. 536, 2016.

[88] A. Cavalcante, A. Mansouri, L. Kacha et al., "Measuring streetscape complexity based on the statistics of local contrast and spatial frequency," *PLoS ONE*, vol. 9, no. 2, Article ID e87097, 2014.

[89] G. Ciocca, S. Corchs, F. Gasparini, E. Bricolo, and R. Tebano, "Does color influence image complexity perception?" in *Proceedings of the International Workshop on Computational Color Imaging*, pp. 139–148, Springer, 2015.

[90] J. Friedenberg and B. Liby, "Perceived beauty of random texture patterns: A preference for complexity," *Acta Psychologica*, vol. 168, pp. 41–49, 2016.

[91] N. Gauvrit, F. Soler-Toscano, and A. Guida, "A preference for some types of complexity comment on "perceived beauty of random texture patterns: a preference for complexity"," *Acta Psychologica*, vol. 174, pp. 48–53, 2017.

[92] C. R. Madan, J. Bayer, M. Gamer, T. B. Lonsdorf, and T. Sommer, "Visual complexity and affect: ratings reflect more than meets the eye," *Frontiers in Psychology*, vol. 8, p. 2368, 2018.

[93] P. Machado, J. Romero, M. Nadal, A. Santos, J. Correia, and A. Carballal, "Computerized measures of visual complexity," *Acta Psychologica*, vol. 160, pp. 43–57, 2015.

[94] G. Ciocca, S. Corchs, and F. Gasparini, "Genetic programming approach to evaluate complexity of texture images," *Journal of Electronic Imaging*, vol. 25, no. 6, Article ID 061408, 2016.

[95] A. Gartus and H. Leder, "Predicting perceived visual complexity of abstract patterns using computational measures: the influence of mirror symmetry on complexity perception," *PLoS ONE*, vol. 12, no. 11, Article ID e0185276, 2017.

[96] D. J. Aks and J. C. Sprott, "Quantifying aesthetic preference for chaotic patterns," *Empirical Studies of the Arts*, vol. 14, no. 1, pp. 1–16, 1996.

[97] R. P. Taylor, A. P. Micolich, and D. Jonas, "Fractal analysis of pollock's drip paintings," *Nature*, vol. 399, no. 6735, p. 422, 1999.

[98] R. P. Taylor, A. P. Micolich, and D. Jonas, "The construction of Jackson Pollock's fractal drip paintings," *Leonardo*, vol. 35, no. 2, pp. 203–207, 2002.

[99] B. Spehar, C. W. G. Clifford, B. R. Newell, and R. P. Taylor, "Universal aesthetic of fractals," *Computers and Graphics*, vol. 27, no. 5, pp. 813–820, 2003.

[100] R. P. Taylor, B. Spehar, C. M. Hagerhall, and P. van Donkelaar, "Perceptual and physiological responses to Jackson Pollock's fractals," *Frontiers in Human Neuroscience*, vol. 5, p. 60, 2011.

[101] K. Jones-Smith and H. Mathur, "Fractal analysis: revisiting Pollock's drip paintings," *Nature*, vol. 444, no. 7119, pp. E9–E10, 2006.

[102] N. Street, A. M. Forsythe, R. Reilly, R. Taylor, and M. S. Helmy, "A complex story: universal preference vs. individual differences shaping aesthetic response to fractals patterns," *Frontiers in Human Neuroscience*, vol. 10, p. 213, 2016.

[103] B. Spehar, N. Walker, and R. P. Taylor, "Taxonomy of individual variations in aesthetic responses to fractal patterns," *Frontiers in Human Neuroscience*, vol. 10, p. 350, 2016.

[104] A. Patuano, *Fractal dimensions of landscape images as predictors of landscape preference [Ph.D. thesis]*, 2018.

[105] C. Viengkham and B. Spehar, "Preference for fractal-scaling properties across synthetic noise images and artworks," *Frontiers in Psychology*, vol. 9, 2018.

[106] G. K. Zipf, *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology*, Addison-Wesley, 1949.

[107] B. Z. Manaris, D. Vaughan, C. Wagner, J. Romero, and R. B. Davis, "Evolutionary music and the Zipf-mandelbrot law: Developing fitness functions for pleasant music," in *Applications of Evolutionary Computing, EvoWorkshop 2003: EvoBIO, EvoCOP, EvoIASP, EvoMUSART, EvoROB, and EvoSTIM*, G. R. Raidl, J.-A. Meyer, M. Middendorf et al., Eds., vol. 2611 of *Lecture Notes in Computer Science*, pp. 522–534, Springer, Essex, UK, 2003.

[108] B. Manaris, J. Romero, P. Machado et al., "Zipf's law, music classification, and aesthetics," *Computer Music Journal*, vol. 29, no. 1, pp. 55–69, 2005.

[109] C. Redies, S. A. Amirshahi, M. Koch, and J. Denzler, "PHOG-derived aesthetic measures applied to color photographs of artworks, natural scenes and objects," in *Proceedings of the European Conference on Computer Vision*, pp. 522–531, Springer, 2012.

[110] N. Lyssenko, C. Redies, and G. U. Hayn-Leichsenring, "Evaluating abstract art: Relation between term usage, subjective ratings, image properties and personality traits," *Frontiers in Psychology*, vol. 7, p. 973, 2016.

[111] S. M. C. Nascimento, J. M. M. Linhares, C. Montagner et al., “The colors of paintings and viewers’ preferences,” *Vision Research*, vol. 130, pp. 76–84, 2017.

[112] H. Weichselbaum, H. Leder, and U. Ansorge, “Implicit and explicit evaluation of visual symmetry as a function of art expertise,” *i-Perception*, vol. 9, no. 2, Article ID 2041669518761464, 2018.

[113] K. Thömmes and R. Hübner, “Instagram likes for architectural photos can be predicted by quantitative balance measures and curvature,” *Frontiers in Psychology*, vol. 9, p. 1050, 2018.

[114] E. E. Rump, “Is there a general factor of preference for complexity?” *Perception & Psychophysics*, vol. 3, no. 5, pp. 346–348, 1968.

[115] A. C. Hall, “Measures of the complexity of random black and white and coloured stimuli,” *Perceptual and Motor Skills*, vol. 29, no. 3, pp. 773-774, 1969.

[116] S. F. Chipman, “Complexity and structure in visual patterns,” *Journal of Experimental Psychology: General*, vol. 106, no. 3, pp. 269–301, 1977.

[117] S. F. Chipman and M. J. Mendelson, “Influence of six types of visual structure on complexity judgments in children and adults,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 5, no. 2, pp. 365–378, 1979.

[118] M. Graves, *Design Judgement Test, Manual*, The Psychological Corporation, New York, NY, USA, 1948.

[119] M. Graves, *The Art of Color and Design*, McGraw-Hill, New York, NY, USA, 1951.

[120] H. J. Eysenck, “Factor analytic study of the maitland graves design judgement test,” *Perceptual and Motor Skills*, vol. 24, pp. 13-14, 1969.

[121] H. J. Eysenck and M. Castle, “Comparative study of artists and nonartists on the maitland graves design judgment test,” *Journal of Applied Psychology*, vol. 55, no. 4, pp. 389–392, 1971.

[122] J. Uduehi, “A cross-cultural assessment of the maitland graves design judgment test using u.s. and nigerian students,” *Visual Arts Research*, vol. 21, no. 2, pp. 11–18, 1995.

[123] K. O. Götz and K. Götz, “The Maitland graves design judgment test judged by 22 experts,” *Perceptual and Motor Skills*, vol. 39, no. 1, pp. 261-262, 1974.

[124] T. Chamorro-Premuzic and A. Furnham, “Art judgment: a measure related to both personality and intelligence?” *Imagination, Cognition and Personality*, vol. 24, pp. 3–25, 2004.

[125] P. Machado and A. Cardoso, “All the truth about NEvAr,” *Applied Intelligence*, vol. 16, no. 2, pp. 101–118, 2002.

[126] J. Romero, P. Machado, A. Carballa, and J. Correia, “Computing aesthetics with image judgement systems,” in *Computers and Creativity*, pp. 295–322, Springer Berlin Heidelberg, Berlin, Heidelberg, Germany, 2012.

[127] G. U. Hayn-Leichsenring, T. Lehmann, and C. Redies, “Subjective ratings of beauty and aesthetics: correlations with statistical image properties in western oil paintings,” *i-Perception*, vol. 8, no. 3, Article ID 2041669517715474, 2017.

[128] J. Braun, S. A. Amirshahi, J. Denzler, and C. Redies, “Statistical image properties of print advertisements, visual artworks and images of architecture,” *Frontiers in Psychology*, vol. 4, p. 808, 2013.

[129] Y. Ke, X. Tang, and F. Jing, “The design of high-level features for photo quality assessment,” in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2006*, vol. 1, pp. 419–426, 2006.

[130] R. Datta, D. Joshi, J. Li, and Z. J. Wang, “Studying aesthetics in photographic images using a computational approach,” in *Proceedings of the Computer Vision – ECCV 2006, 9th European Conference on Computer Vision, Part III, LNCS*, pp. 288–301, Springer, Graz, Austria, 2006.

[131] R. Datta, D. Joshi, J. Li, and J. Z. Wang, “Image retrieval: ideas, influences, and trends of the new age,” *ACM Computing Surveys*, vol. 40, no. 2, pp. 5:1–5:60, 2008.

[132] S. Hacker and L. von Ahn, “Matchin: eliciting user preferences with an online game,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI ’09*, pp. 1207–1216, Boston, MA, USA, April 2009.

[133] A. Carballal, A. Santos, J. Romero, P. Machado, J. Correia, and L. Castro, “Distinguishing paintings from photographs by complexity estimates,” *Neural Computing and Applications*, vol. 30, no. 6, pp. 1957–1969, 2018.

[134] F. Lemarchand, “Fundamental visual features for aesthetic classification of photographs across datasets,” *Pattern Recognition Letters*, vol. 112, pp. 9–17, 2018.

[135] J. McCormack, “Working with generative systems: an artistic perspective,” in *Proceedings of the Electronic Visualisation and the Arts (EVA 2017), Electronic Workshops in Computing (eWiC)*, J. Bowen, N. Lambert, and G. Diprose, Eds., pp. 213–218, BCS Learning and Development Ldt., London, UK, 10–13 July 2017.

[136] P. Machado, J. Romero, and B. Manaris, “Experiments in computational aesthetics: an iterative approach to stylistic change in evolutionary art,” in *The Art of Artificial Evolution: A Handbook on Evolutionary Art and Music*, J. Romero and P. Machado, Eds., pp. 381–415, Springer Berlin Heidelberg, 2008.

Hindawi
Complexity
Volume 2019, Article ID 4659809, 12 pages
https://doi.org/10.1155/2019/4659809

Research Article

Avoiding the Inherent Limitations in Datasets Used for Measuring Aesthetics When Using a Machine Learning Approach

Adrian Carballal,[1] **Carlos Fernandez-Lozano**,[1,2] **Nereida Rodriguez-Fernandez**,[3] **Luz Castro**,[3] **and Antonino Santos**[1]

[1] *Computer Science Department, Faculty of Computer Science, University of A Coruña, A Coruña 15071, Spain*
[2] *Investigación Biomédica de A Coruña (INIBIC), Complexo Hospitalario Universitario de A Coruña (CHUAC), A Coruña 15006, Spain*
[3] *Computer Science Department, Faculty of Communication Science, University of A Coruña, A Coruña 15071, Spain*

Correspondence should be addressed to Adrian Carballal; adrian.carballal@udc.es

Received 17 August 2018; Revised 14 November 2018; Accepted 20 December 2018; Published 8 January 2019

Guest Editor: Jon McCormack

An important topic in evolutionary art is the development of systems that can mimic the aesthetics decisions made by human begins, e.g., fitness evaluations made by humans using interactive evolution in generative art. This paper focuses on the analysis of several datasets used for aesthetic prediction based on ratings from photography websites and psychological experiments. Since these datasets present problems, we proposed a new dataset that is a subset of DPChallenge.com. Subsequently, three different evaluation methods were considered, one derived from the ratings available at DPChallenge.com and two obtained under experimental conditions related to the aesthetics and quality of images. We observed different criteria in the DPChallenge.com ratings, which had more to do with the photographic quality than with the aesthetic value. Finally, we explored learning systems other than state-of-the-art ones, in order to predict these three values. The obtained results were similar to those using state-of-the-art procedures.

1. Introduction

Estimating aesthetic value and the complexity of an image is a technological challenge that has recently been addressed by numerous fields, including psychology and artificial intelligence. Several research groups have attempted to create computer systems that are able to learn the aesthetics perception of a group of human beings as a part of a generative system (such as evolutionary art systems) or that can be used for automatic image selection or ordering. Given the subjective nature of the aesthetic problem, the selection of the dataset for the training is vital. This paper explores a new way to build a dataset and provide initial results by using machine learning techniques.

Previous research studies [1, 2] have concluded that the degree of generalisation of some existing sample sets was not enough to take them as reference in the training of automated prediction and classification of images. Other functional limitations were identified in these datasets, which are also mentioned in this paper.

In order to solve the problems identified in these datasets, this paper describes the creation of a new set of images from the website DPChallenge.com, with greater statistical consistency. Besides, this new dataset was evaluated in terms of aesthetics and quality by a group of individuals under controlled experimental conditions. This makes it the first dataset evaluated by two different populations (the one evaluating at the DPChallenge.com portal and the one evaluating it in person).

With the new dataset created, several Machine Learning-based models were trained for the automated prediction of the aesthetic and quality value and that of DPChallenge.com.

This paper starts with a state-of-the-art section on the datasets created for the automated prediction and classification of images. In Section 3, the limitations found in such sample sets are provided. Section 4 describes the method for the creation of a new dataset with greater statistical coherence and the results of the evaluation procedure obtained under experimental conditions. Section 5 presents the Machine Learning models based on the prediction that were used as well as the results obtained in the training based on the three available criteria for the images of the proposed set. There is a section discussing the results and another one with the final conclusions.

2. State of the Art

Some authors, such as Datta et al. [3], Wang et al. [4], Ke et al. [5], and Luo et al. [6], conducted studies aimed at automated aesthetic classification using a number of technical characteristics such as lightness, saturation, Rule of Thirds, etc. For these experiments, sets of large-format photographs from websites and the evaluations made by the users of such sites were used. On the other hand, other authors, including Cela-Conde et al. [7], Forsythe et al. [8], and Nadal et al. [9], carried out aesthetic perception and image complexity experiments using a sample set with a more limited number of images, but evaluated by a specific set of people under controlled experimental conditions. A brief analysis of these sample sets is presented below.

2.1. Photo.net (2006). Datta et al. [3] created a dataset based on the photography website Photo.net, which has over a million images and 400,000 users. In this dataset, each image is rated on a range from 1 to 7 (1 being the worst possible score and 7 the best) based on aesthetics and originality. Statistical information on the rating can be found on the website. It does not provide information on the image evaluators, though. The full dataset comprises 3,581 images rated by at least 2 persons and has an average score between 3.55 and 7 and an overall total average of 5.06, with a standard deviation of 0.83. The high correlation found between the criterion of originality and aesthetics (Pearson's r = 0.891) might indicate that users most assuredly are not making such distinctions.

Datta et al. [3] and other researchers such as Wong et al. [8], who used this sample group, have established a division to obtain two different groups: (i) the images with an average score equal to or higher than 5.8 were branded high quality and (ii) those with scores equal to or lower than 4.2 were branded low quality. In the case of the study conducted by Datta et al. [3] a success rate of 70.12% was achieved in the global classification using Support Vector Machines (SVM): 68.08% for high quality images and 72.31% for low quality images.

2.2. Photo.net (2008). In 2008, a new study was published by Datta et al. [10], which introduced a second set of data from the website consisting of 20,278 images rated by an average of 16.81 persons with a standard deviation of 16.19. It should be noted that there were images evaluated by a minimum of four people and others by a maximum of 395. When comparing this study with the previous one, it becomes apparent that this statistical analysis is more complete, as it provides specific data for each image. The total set of images had at least four ratings per image, with scores ranging between 2.33 and 6.99, and a global average of 5.15, with a mean standard deviation of 0.58. From the same set, Wong et al. [8] displayed 44 metrics grouped into three categories with global characteristics, for which they used a reduced set of images from the original experiment down to a total of 3,161. After performing a classification using SVM with linear kernel and resorting to a crossed validation with 5 independent runs, 78.2% of the images were successfully classified (82.9% high quality and 75.6% low quality).

2.3. DPChallenge.com. Ke et al. [5] created a different sample set, which became one of the most commonly used in aesthetic classification experiments. For the construction of this set, the photography portal DPChallenge.com was used, with a total of 60,000 images rated by at least 100 persons being selected.

For the aesthetic classification experiments, two sets of 6,000 photographs were created by selecting the top and bottom 10% after arranging them according to their mean score. Subsequently, Ke et al. [5] carried out a subdivision into two new random subsets, thus obtaining 4 sets of 3,000 images (two high quality and two low quality sets). A set of each type was used to train the proposed systems, while the other was used to validate their capacity and efficacy.

2.4. Dataset Created by Psychologists. Cela-Conde et al. [7] created a dataset consisting of a final standardized set of 800 images divided into 5 categories: artistic abstract (AA), non-artistic abstract (AN), artistic representational (RA), non-artistic representational (RN), and photographs of natural scenes and human constructions (NHS).

The images were shown to a group of 240 participants (112 men and 128 women, with a mean age of 22.03 years and a standard deviation of 3.75), randomly divided into subgroups of 30 persons in a controlled experimental environment. The images were shown for five seconds and participants were asked to rate the visual complexity of a subset of stimuli on a Likert scale from 1 to 5 (1 being the worst possible score and 5 the best). Consequently, each image had a total of 30 ratings. The mean value obtained by each subgroup for each stimulus was the value considered to represent the complexity of this stimulus in the final set. The stimuli in this set were used by Cela-Conde et al. [7], Forsythe et al. [8], Nadal et al. [9], and Machado et al. [11].

3. Limitations Found in the Dataset Available

The study of the generalisation capacity of the analysed datasets led to the conclusion that they did not provide a satisfactory degree of generalisation: the correlation is greater when the validation set belongs to the same source of data as the training set. However, in experiments where the test was performed with a set from a source different from the training set, the correlation results decreased notably. A clear example in this regard can be seen in experiments conducted in previous research studies [1, 2]: when training a subset of 6,000 images from DPChallenge.com carried out by Ke et al. [5], the result of the correlation was 91.38%. If validated with another subset from the same source, however, the resulting percentage decreased down to 56.21%, when using, for example, the dataset from Photo.net created by Datta et al. in 2006 [3], and down to 55.39% with the dataset from Photo.net created by Datta et al. in 2008 [10].

Besides, the sample sets trained with ratings from the photography portals had some defects: the evaluation system did not have the same control as a psychological test because it was not possible to obtain all the information about the evaluating users or about the device used to see the image (smartphone, computer), or distance or lighting conditions; the amount of images might be insufficient as there was no justified reason to choose a sample size and there was a huge difference in the number of people rating each image; user evaluations could be easily conditioned by personal relationships with the creator of the work or a momentary surge in popularity of certain styles. Lastly, in one of the cases [3] it was shown that the users of these portals did not have enough basis to differentiate between aesthetic and originality criteria, with Pearson's correlation coefficient of 0.891. Furthermore, as these datasets were designed for binary classification, only the images rated with extreme scores (those obtaining the highest and lowest ratings) were used, leaving out of the set the images with intermediate ratings.

In the set created by Ke et al. [5] there was another limitation in the collected evaluations, as the DPChallenge.com portal operated as if it were a photography competition and there was no specification of any criteria to assess the images. Consequently, any user can evaluate the image on their own criteria, which may have nothing to do with those of other people.

On the other hand, in the dataset created by Cela-Conde et al. [7] the number of images presented by category was not balanced. Therefore, the obtained results cannot be considered as representative of the whole. Besides, the set was built on the basis of a considerable number of subsets of images, which resulted in the dataset eventually becoming a number of datasets of independent themes of smaller size, with less internal coherence.

Once the limitations of the studied datasets were identified, a new dataset was built for the aesthetic prediction of images. This dataset was evaluated by humans under controlled experimental conditions using a coherent set of images.

4. Building a New Dataset

After identifying the limitations discussed above in the existing sets of images, we created a new dataset for the prediction and classification of images, in which the process of human evaluation of the images was carried out under controlled experimental conditions. This new method is generally put forward in [1] and includes the advantages of the sets of images studied in this article. This new method of creation makes it possible to build a set of images with greater statistical coherence from the rating results on the photography website DPChallenge.com and is subsequently evaluated in a manner similar to the procedure used by Forsythe et al. [8]. Thus, we shall be able to analyse the correlation between the results obtained with subjects under controlled circumstances and those obtained through the photography portal.

4.1. Source Data. We began by collecting a set of images from the DPChallenge.com photography portal. The images on the DPChallenge.com portal are rated by users within the range [1, 10], where 1 is the lowest possible score and 10 the highest. The only information about the score in DPChallenge.com is that a score of 1 is a "bad" photo, and a score of 10 is a "good" photo. So the score is not clearly related to aesthetics, photographic quality, or originality. Nevertheless, this portal has been used in the past to obtain data for aesthetic classification experiments [5, 12, 13]. The original idea behind this site was for it to be a place where friends could teach themselves to be better photographers by giving each other a "challenge" for the week. Methodologically, DPChallenge organises weekly competitions into "themes" represented by a word of phrase (e.g., "Alfred Hitchcock", "Abstract: Black and White II", "Color Portrait IV"). For the current study, this aspect of the evaluation is not taken into account.

Images were collected using a brute force process whereby all data from all images whose identifiers were between 10,000 and 172,000 in May 2012. All statistical information of the ratings was available for only 40,047 images. The images in this initial set were rated by an average of 233 subjects and the mean rating was 5.23 ± 0.78. All descriptive data are shown in Figure 1(a). The file with the evaluation data and the links to the images used (for copyright reasons) are publicly available at https://doi.org/10.6084/m9.figshare.6127295.v1. Figure 1(c) shows the arrangement of votes based on each range and Figure 1(b) displays the distribution of the mean evaluations of the images within the range of scores, showing in both cases that they apparently follow a Gaussian model.

4.2. Dataset Proposed. As noted above, only the images in which all the evaluation data were available were used. Then, only the images with at least 100 ratings were selected. The objective was that the mean value subsequently attributed to each image was the least biased possible.

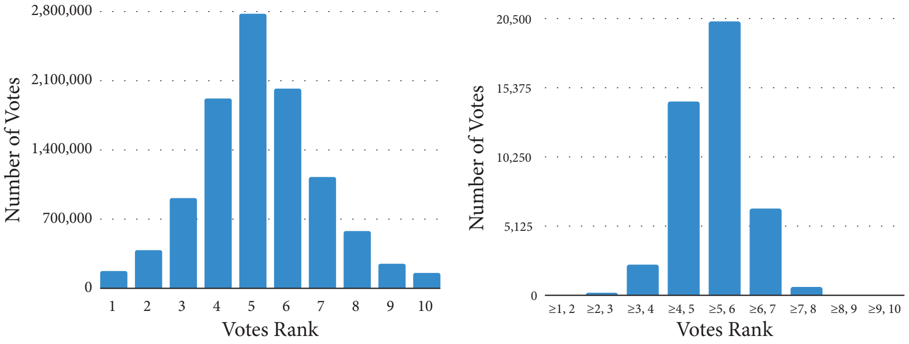
Once this selection was made, images were arranged in groups according to the mean ratings given on DPChallenge.com. The images in our selection were classified according to 9 scoring ranges, one for each integer value of valid evaluation. Then, a minimum number of images were set for all groups. In our case, the minimum number was 200 (see Figure 2(b)). There were no sets of images numerous enough with mean scores below 3 or higher than 8. Consequently, the used groups were collected from the [3, 8] range. From these groups, 200 images with the lowest standard deviation were selected. In other words, these were images with the most internally consistent scores. We used the more consistent image set in order to build a dataset that can be used as ground truth dataset. The descriptive data for each of the ranges are detailed in Table 1. Figure 2 shows (a) the distribution of the number of votes within the range of valid

Table 1: Descriptive data for each of the five sets of 200 images that make up the proposed dataset.

Range	[3, 4)	[4, 5)	[5, 6)	[6, 7)	[7, 8)
Average	3.5943	4.4695	5.4975	6.4715	7.3112
Deviation	0.2613	0.2868	0.2894	0.2845	0.2335
Variance	0.0683	0.0822	0.0837	0.0809	0.0545
Kurtosis	-0.8224	-1.2370	-1.1765	-1.1595	-0.4500
Bias	-0.3998	0.1611	-0.0005	0.1099	0.6879
Minimum	3.0070	4.0130	5.0060	6.0030	7.0000
Maximum	3.9970	4.9970	5.9970	6.9940	7.9530

Descriptive data	
Images	44047
Average	5.2405
Deviation	0.7821
Variance	0.6117
Kurtosis	0.2480
Bias	-0.0182
Minimum	1.9951
Maximum	8.3900

(a)



(b) (c)

Figure 1: Characterisation of all 44,047 images initially obtained from DPChallenge. (a) Descriptive data, (b) arrangement of the number of votes within the range of valid ratings, and (c) distribution of mean image evaluations within the range of valid ratings.

scores and (b) the distribution of the mean ratings within the range of valid scores for the 1000-image dataset.

This process provides a set of images with equal number of elements for each range, with high scoring consistence, and which could eventually be the most representative.

4.3. Human Evaluation. The dataset proposed above was evaluated by a number of humans under controlled experimental conditions. According to Infinite Population Sampling [14] with a minimum sample size of 8 individuals and 95% of confidence level, the true population rating of an image can be obtained, with a margin of error of 3%.

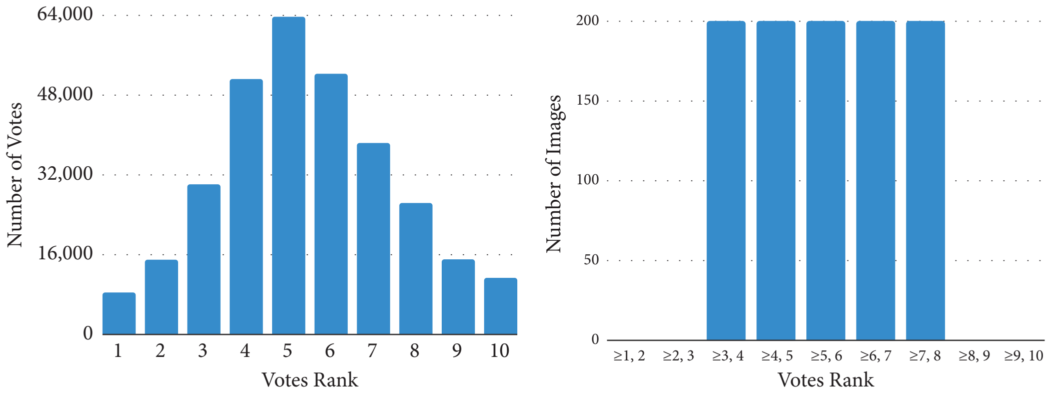
To this end, 5 subsets were created with randomly selected images out of a total of 1,000 available. Each person could rate the images in one or several of these subsets with a score between 1 and 5, where 1 is the lowest possible score and 5 the highest. Each set was evaluated by at least 10 persons (a total of 10,000 ratings).

Evaluations were carried out on February 1st and March 5th, 2018, by student volunteers of the University of A Coruña, Spain (mainly, students at the School of Communication Sciences). Ninety (33 male and 66 female) participants (18.7 years, age range 18-30) took part in this study. Each participant evaluated at least 200 images before the research study and under the same viewing conditions: screens with the same specifications, same lighting conditions, and same distance between evaluators and the screens.

For every image, users independently rated its aesthetic value and quality. The English translation of the text of the survey questions verbatim is: "In this task we want you to evaluate the quality and aesthetic value of each of the images that we propose. To score the "quality" you should look at the framing, focus, colors, etc. In general, professional photographs have higher quality than photographs taken by amateurs. The editing of images (use of Photoshop, filters, etc.) does not have to affect its quality. It may be that you do not like an image, but if it is well made, your quality score should be high. For the aesthetic score value we look for your personal opinion about the image, whether you like it or not. The semantic value should not influence. That is, a nice picture of a crying baby can have a high aesthetic value score."

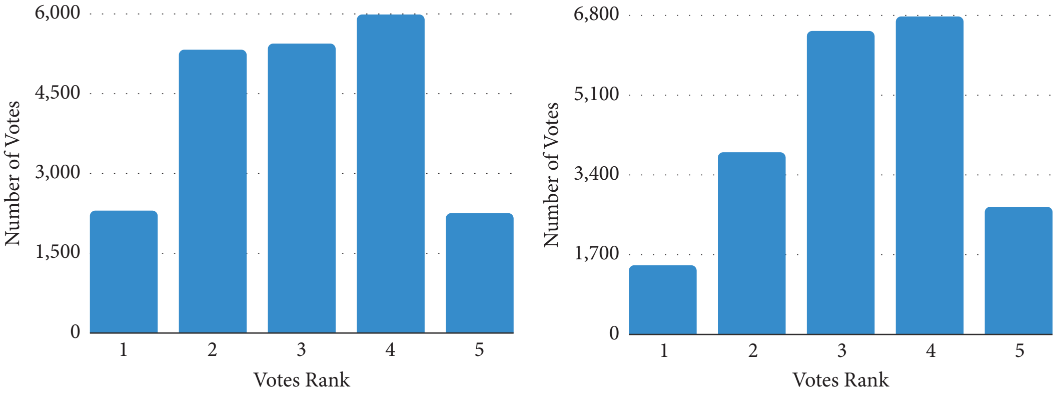
The data shown in Figure 3 correspond to the mean obtained for each image from the different evaluations made for both aesthetic and quality criteria.

The correlation between the scores given in person and those registered on the Dpchallenge.com platform was calculated (see Figure 4). Pearson's correlation between the mean score on Dpchallenge.com and the mean score was 0.692 according to the aesthetic criterion and 0.690 according to Spearman's. The mean correlation between DPChallenge.com and the mean according to the quality value was 0.748 according to Pearson's and 0.756 according to Spearman's. Lastly, the correlation between the two measures obtained in the in-person experiment (aesthetics/quality) was 0.787 according to Pearson's and 0.786 according to Spearman's, higher than in the other two correlations. Figure 4 shows the Scatterplots between ranks for the three possible combinations given the three criteria that are evaluated for the entire study.



(a) (b)

Figure 2: Characterisation of the 1000 images in the proposed set. (a) Distribution of the number of votes within the scoring range and (b) distribution of mean ratings in the images within the valid range of scores.



(a) (b)

Figure 3: Distribution of the mean aesthetic (a) and quality (b) ratings obtained in the control group.

5. Machine Learning Approach

In this study, some state-of-the-art models based on Machine Learning applied to the proposed input were proposed. The aim of these experiments was to study whether the existing correlation values between both human populations seen in the previous sections (DPChallenge and control group) can be *replicated* by a computer system for the proposed dataset.

5.1. Materials and Methods. To characterise the images that make up the study set, a feature extractor available in WND-CHARM [15] was used, which is a multipurpose image classifier that can be applied to a wide variety of image tasks. According to its developers, the system extracts a large set of image features, including polynomial decompositions, high contrast features, pixel statistics, and textures, among others. These features are computed on the raw image, transforms of the image, and transforms of transforms of the image. The final feature vector comprises 2905 variables, each of which reporting on a different aspect of image content. All features are based on greyscale images, so colour information is not currently used.

The authors tested the different computational models using a 10-fold cross-validation to split the data and 50 runs per model in order to evaluate the performance across different experiments. The performance of the models was evaluated using Spearman's correlation coefficient (rho) and Pearson's correlation coefficient (Pearson's r).

5.2. Computational Models. The authors performed several experiments in order to select the best model using the R package and MATLAB©. Some of the used computational models looked for the smallest subset of variables of the original set which provided a better performance [16], or at least equal to that obtained when using all the possible variables, considering this was a Feature Selection (FS) approach [17–19].

More specifically, the used methods were the following: the well-known Support Vector Machines-Recursive Feature Elimination (SVM-RFE) [20, 21] and the Generalized Linear Model with Stepwise Feature Selection (GLM) [22] which selects features that minimise the AIC score and the most basic standard Multiple Linear Regression (LM) without FS. The abilities of the RRegrs Package [23] were enhanced in

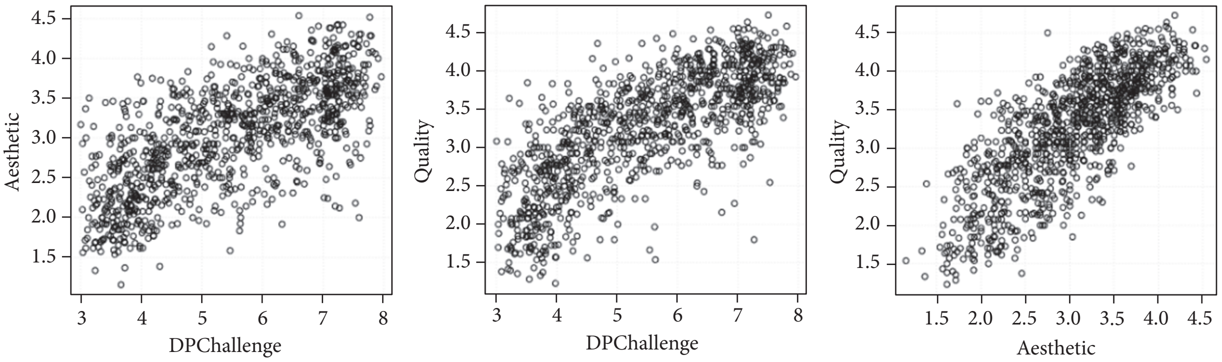


Figure 4: Scatterplots between ranks for the three possible combinations given the criteria evaluated for the entire study.

order to implement the SVM-RFE and GLM to avoid finding the best model according to the proposed methodology as, according to [24], it should be performed based on a null hypothesis test. This package was also enhanced in order to avoid the initial splitting process, and an external cross-validation process was performed to avoid selection bias as suggested by [25]. The last step was modified in order to easily extract the results for all the models.

The K-nearest neighbour (k-NN) algorithm is a technique based on the cluster theory. In this case, a variant called weighted k-NN [26] was used. It is based on the fact that a new observation particularly close to an observation within the learning set should have great importance in the decision-making process and, conversely, an observation that is at a further distance should have much less importance [27]. For this algorithm, only the hyperparameter k was tuned, which represented the number of neighbour data points that were considered closest. The range of values was from 1 to 5.

The generalized boosted models (GBM) applied the approach described in [28], to establish the foundation of boosting algorithms. GBM estimation involves an iterative process with multiple regression trees to capture complex and nonlinear relationships without overfitting the data [29, 30]. It works with continuous and discrete variables and is invariant to their monotonic transformations [31]. For this algorithm, the interactive depth was represented by the number of splits it had to perform on a tree (starting from a single node) and the number of trees that were tuned. The range of values used was from 1 to 4 and 100, 250, and 500 for the number of trees.

The design of our experiments was based on a novel methodology for the development of experimental designs in regression problems with multiple machine learning regression algorithms [32]. For each model described above, the optimal set of parameters was sought using hyperparameter optimisation.

5.3. Results. Figure 5 and Table 2 show the results obtained for each of the four methods studied according to Pearson's and Spearman's correlation value, using as reference the average ratings from the DPChallenge.com portal. Firstly, examining Spearman's correlation values, the maximum value of the SVM-based model was 0.574, using 1024 variables. The input set could be decreased down to 256 with no significant loss of performance (0.570), as the correlation values remained statistically constant between both figures. On the other hand, if we look at the values for Pearson's r, the same pattern remained, since, with 1024 input variables, 0.581 was obtained, whereas, with 256, 0.574 was obtained (with no significant difference in performance). In any case, both Spearman's and Pearson's values show a moderate uphill (positive) relationship, with the exception of k-NN.

The authors checked the significance of the difference between GLM, SVM, GBM, and k-NN with 256 input variables (see Figure 6) using a Kruskal-Wallis test, and our results showed that, with a very high level of confidence, SVM (cost = 2^{-6} y gamma=2^{-9}) was significantly better than the others with a p-value < 2.2 x 10^{-16}. Consequently, it could be stated that the minimum input set with the best results was the one with 256 input variables in combination with an SVM prediction model with specified parameters.

Once the method with the best results was identified using the average ratings obtained by the users of DPChallenge.com, the best SVM hyperparameters were calculated (cost = 2^{-4} and gamma=2^{-12} in both cases) training the scores for "aesthetics" and "quality" obtained in the above-mentioned experiment with humans.

As shown in Figure 7, the values for any of the 3 cases are below 0.60 on average. Specifically, it was 0.578 for DPChallenge, 0.456 for aesthetics, and 0.539 for quality, using as mean of performance Spearman's rho and 0.574, 0.451, and 0.562, respectively, using Pearson's r. On the negative side, it is particularly relevant that in the case of "aesthetics" there is a weak uphill (positive) relationship given the average value obtained with both measures.

6. Discussion

A correlation of 0.78 was obtained between the ratings based on aesthetics and those based on quality. This indicated that the evaluation teams distinguished between both criteria when compared with the measurements made by Datta et al. [3], where Pearson's correlation between aesthetics and originality was 0.891.

Table 2: Average results presented in Figure 5, identifying hyperparameters and input size for each model.

Size	Model	Pearson	SD	Hyperparameters
16	GLMNET	0.5320	0.0717	Alpha=0
16	GBM	0.5234	0.0738	Interaction.depth=4, n.trees=500
16	k-NN	0.4831	0.0744	k=12; distance=2
16	SVM	0.5389	0.0713	Cost = 16 Gamma=0.00984
32	GLMNET	0.5451	0.0709	Alpha=0
32	GBM	0.5266	0.0733	Interaction.depth=4, n.trees=500
32	k-NN	0.4851	0.0750	k=12; distance=2
32	SVM	0.5581	0.0732	Cost=0.397 Gamma=0.00984
64	GLMNET	0.5406	0.0669	Alpha=0
64	GBM	0.5474	0.0723	Interaction.depth=4, n.trees=500
64	k-NN	0.4898	0.0752	k=12; distance=2
64	SVM	0.5503	0.0691	Cost=2.52 Gamma=0.000244
128	GLMNET	0.5473	0.0745	Alpha=0
128	GBM	0.5425	0.0679	Interaction.depth=4, n.trees=500
128	k-NN	0.4926	0.0720	k=12; distance=2
128	SVM	0.5687	0.0676	Cost=0.397 Gamma=0.00155
256	GLMNET	0.5555	0.0719	Alpha=0,15
256	GBM	0.5479	0.0704	Interaction.depth=4, n.trees=500
256	k-NN	0.4776	0.0774	k=12; distance=2
256	SVM	0.5778	0.0671	Cost=2.52 Gamma=0.000244
512	GLMNET	0.5748	0.0701	Alpha=0,15
512	GBM	0.5482	0.0765	Interaction.depth=4, n.trees=500
512	k-NN	0.4845	0.0758	k=12; distance=2
512	SVM	0.5747	0.0683	Cost=2.52 Gamma=0.000244
1024	GLMNET	0.5644	0.0708	Alpha=0,15
1024	GBM	0.5473	0.0685	Interaction.depth=4, n.trees=500
1024	k-NN	0.4908	0.0777	k=12; distance=2
1024	SVM	0.5782	0.0670	Cost=0.397 Gamma=0.000244
2048	GLMNET	0.5602	0.0733	Alpha=0,15
2048	GBM	0.5465	0.0692	Interaction.depth=4, n.trees=500
2048	k-NN	0.4482	0.0815	k=12; distance=2
2048	SVM	0.5723	0.0685	Cost = 2.52 Gamma=0.000244
fulldataset	GLMNET	0.5590	0.0719	Alpha=0,15
fulldataset	GBM	0.5476	0.0690	Interaction.depth=4, n.trees=500
fulldataset	k-NN	0.4299	0.0825	k=12; distance=2
fulldataset	SVM	0.5554	0.0721	Cost=2.52 Gamma=0.000244

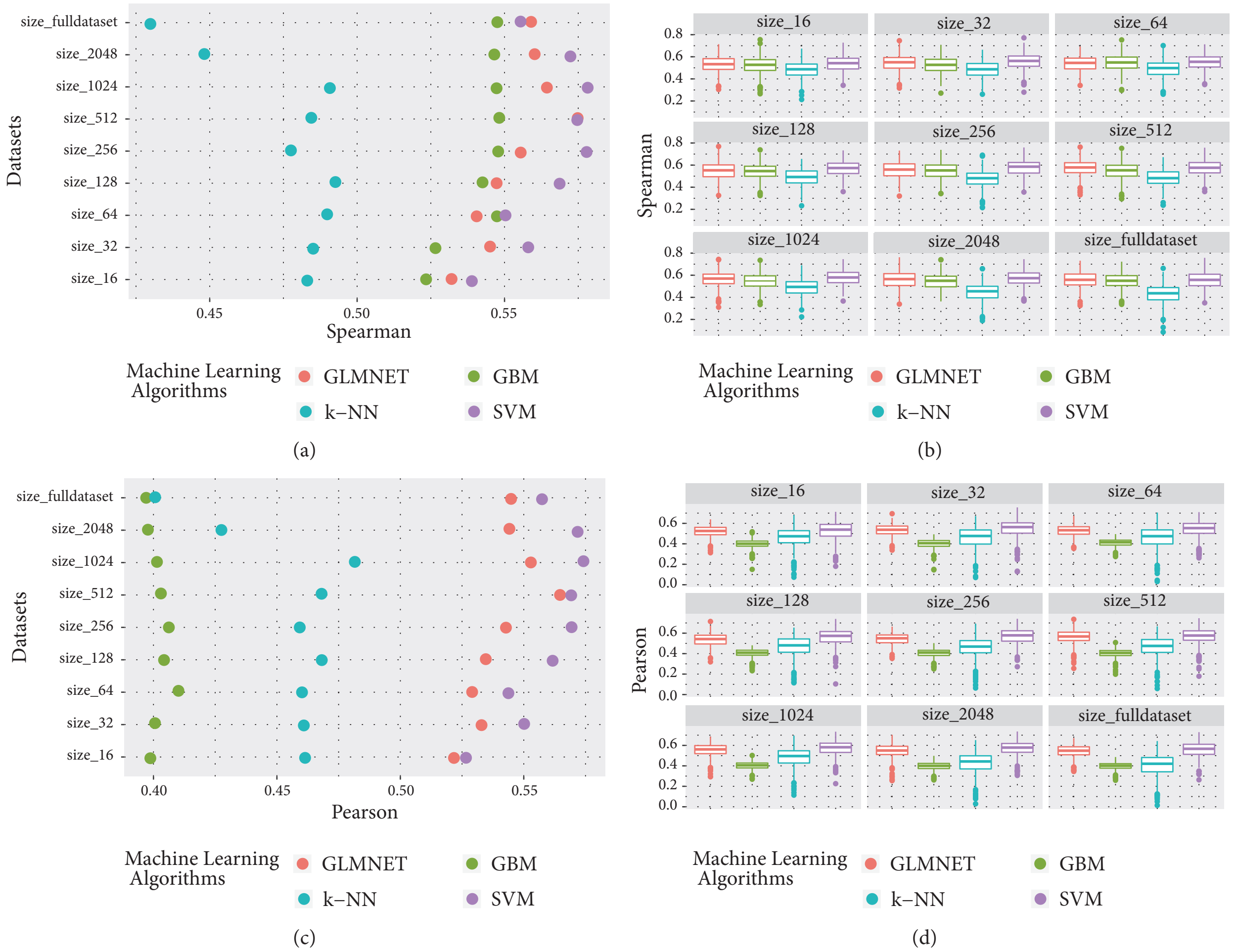
Regarding the correlation between DPChallenge and quality and aesthetics individually, we should begin by underscoring that the highest correlation was between DPChallenge and quality, which suggests that, at DPChallenge, the photographic quality is valued over the aesthetic value of the image.

In our opinion, there was no single reason that explained the difference between the correlations regarding DPChallenge, as far as the aesthetic and quality values were concerned:

(i) In the case of DPChallenge, the users' rating may be conditioned by affinity with the author of the photograph as we were dealing with a competition whereas in the case of the control group, the experimental conditions were controlled (for instance, everyone used the same screen model, at the same distance, with the same ambient light, etc.).

(ii) At DPChallenge, numerous devices can be used (smartphones, tablets, and high resolution screens) and conditions such as viewing distance and ambient light are heterogeneous.

(iii) In the case of the in-person group, the evaluation criteria were established: aesthetics and quality. At DPChallenge, as mentioned above, we were dealing with a photography competition and many different things may be evaluated such as quality, aesthetics, originality, etc.



(a)

(b)

(c)

(d)

Figure 5: Results obtained for the four models proposed and optimised by hyperparameterisation. On the right, the mean values for Spearman's (top) and Pearson's r (bottom) are shown. On the left, the distributions of all 50 independent runs for each optimum model (Spearman top and Pearson bottom) are shown with different input sizes tested using FS.

(iv) On the other hand, in the case of in-person ratings, the minimum per image was 10 whereas, for the evaluations from DPChallenge, the minimum was 100 for each image. It should be borne in mind that the used images had the lowest standard deviation at DPChallenge, which means that the mean rating at DPChallenge had a standard deviation (0.27) lower than that of in-person ratings (1.18 for aesthetics and 1.10 for quality).

If we pay attention to the visual characteristics of some of the images of the set (Figure 8), some noteworthy cases were found:

(i) Figure 8(a) was wrongly rated by the users of the photography portal as having some overexposed areas. It showed a palm tree on the foreground which was slightly incorrectly exposed. However, it obtained a high score in aesthetics because it had some aesthetic value for the evaluators (these motifs tend to have certain aesthetic value). The value of quality was closer to that of DPChallenge in this case.

(ii) Figure 8(b) in DPChallenge obtained a low rating, whereas as far as quality and aesthetics were concerned, it was clearly over average. This difference could be due to the experimental conditions in which the in-person evaluation took place (good quality of image on a big-enough screen, well exposed sky). Under these conditions, evaluators might have paid more attention to the drop and the sky to the detriment of darker area.

(iii) Figure 8(c) at DPChallenge had a high rating, which may be due to the fact that its originality and editing were taken into account.

(iv) Lastly, in Figure 8(d) quality was again closer to DPChallenge. However, a lower score was given in aesthetics.

All this shows that, at DPChallenge, in specific cases, different parameters might be evaluated: originality, quality, aesthetics, photo editing, etc.

As to the use of machine learning techniques to predict each of the three criteria studied, the highest correlation

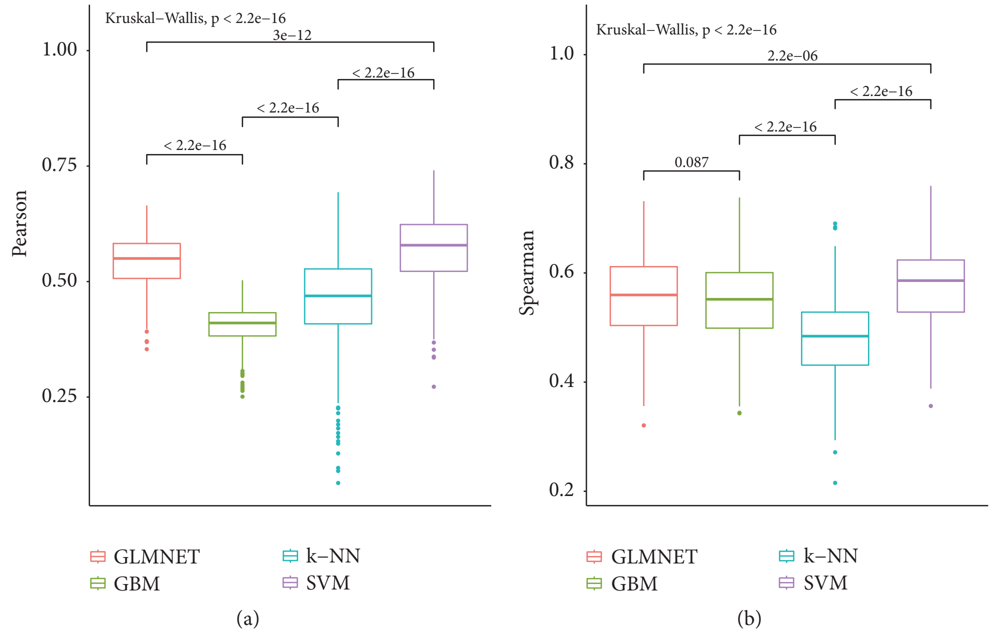


FIGURE 6: Distribution of the correlations obtained for each optimised model (Pearson's on the right and Spearman's on the left). For each pair, the p-value obtained using a Kruskal-Wallis test is shown.

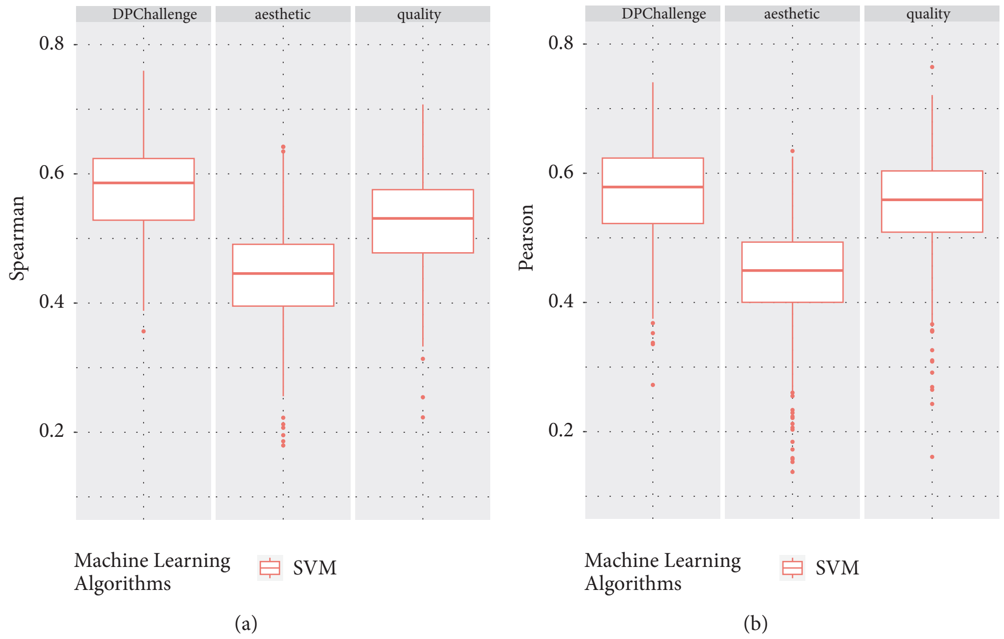


FIGURE 7: Distribution of correlations (Spearman's on the right and Pearson's on the left) obtained for each of the three criteria (DPChallenge, Aesthetic, and Quality) using 256 input variables and an SVM model optimised using hyperparameterisation.

obtained was 0.578 using SVM. This value is similar to those obtained by Marin and Leder [33] using as criteria "arousal" (Spearman's rho=0.44) and "pleasantness" (Spearman's rho=0.64) or by Nadal [9] with "beauty" (Spearman's rho=0.648) under similar experimental conditions with humans. These values were obtained using numerous state-of-the-art methods in predicting and determining the best configuration for each of them through hyperparameterisation.

As to the correlation between the SVM model with quality and aesthetics individually (Figure 8), it follows that for the system it was simpler to learn the quality values than

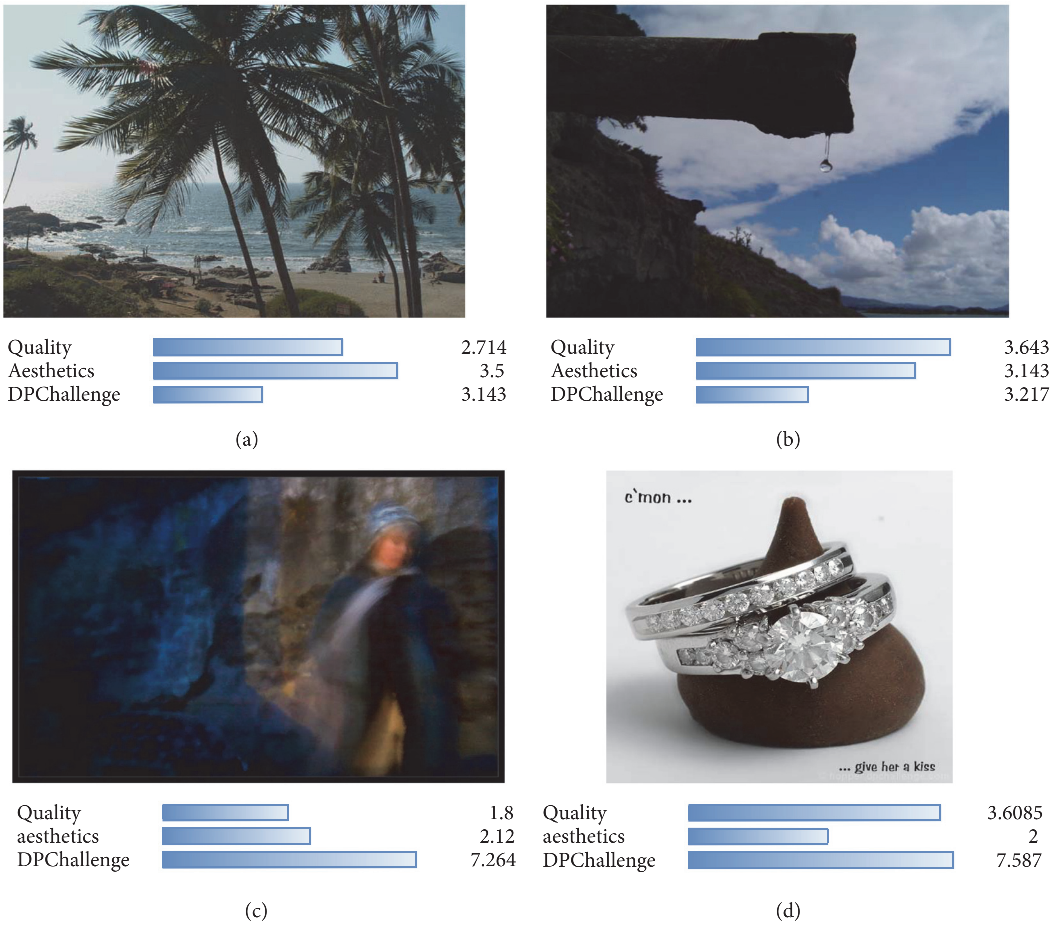


Figure 8: Examples of images with different scores based on the three evaluation criteria. For each image, a number value is given according to each criterion, while bars show the normalised weight of such value within each assessment range (DPChallenge in the [1, 10] range and aesthetics and quality in the [1, 5] range).

the aesthetic ones, which makes sense considering that the former is a less subjective component and more related to the characteristics of the image.

7. Conclusions

Taking into account a number of problems found regarding the state-of-the-art datasets, a dataset was developed following a new methodology. This dataset consists of 1000 images from the DPChallenge portal, which were evaluated in 3 different ways: (1) evaluation from the DPChallenge portal with at least 100 scores per image; (2) an aesthetic evaluation conducted under controlled experimental conditions and a minimum of 10 votes per image; (3) a quality assessment made under the same conditions as (2). As far as the authors are aware, this is the first time a dataset is evaluated based on three different criteria by two different populations.

The results of the correlation suggest that the evaluation of DPChallenge is closer to a quality criterion than to an aesthetic one. The DPChallenge users and in-person evaluators rate images differently and it is apparent that at DPChallenge each user may be following different criteria for the evaluation of images, such as originality, image editing, quality, aesthetics, etc.

Numerous state-of-the-art computational techniques were used and their optimal configurations were identified and applied to all three criteria (DPChallenge, aesthetic, and quality) and correlations of 0.578, 0.456, and 0.539, respectively, were achieved. These results are similar to those obtained in the state-of-the-art experiments. They show that machine learning techniques are more able to learn human assessment of technical quality than aesthetic value, despite the fact that the gap between them is very narrow.

It should be emphasized that machine learning approaches are better at predicting quality than aesthetics, perhaps because of their lower subjective component and their greater association with the intrinsic characteristics of the images.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by the General Directorate of Culture, Education and University Management of Xunta de Galicia (Ref. GRC2014/049) and the European Fund for Regional Development (FEDER) allocated by the European Union, the Portuguese Foundation for Science and Technology for the development of project SBIRC (Ref. PTDC/EIA–EIA/115667/2009), Xunta de Galicia (Ref. XUGA-PGIDIT-10TIC105008-PR) and the Spanish Ministry for Science and Technology (Ref. TIN2008-06562/TIN), and the Juan de la Cierva fellowship programme by the Spanish Ministry of Economy and Competitiveness (Carlos Fernandez-Lozano, Ref. FJCI-2015-26071). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

[1] A. Carballal, L. Castro, N. Rodríguez-Fernández, I. Santos, A. Santos, and J. Romero, "Approach to minimize bias on aesthetic image datasets," in *Interface Support for Creativity, Productivity, and Expression in Computer Graphics*, p. 131, IGI Global, 2019.

[2] A. Carballal, L. Castro, R. Perez, and J. Correia, "Detecting bias on aesthetic image datasets," *International Journal of Creative Interfaces and Computer Graphics*, vol. 5, pp. 62–74, 2014.

[3] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Computer Vision – ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds., pp. 288–301, Springer, Berlin, Germany, 2006.

[4] W. Wang, D. Cai, L. Wang, Q. Huang, X. Xu, and X. Li, "Synthesized computational aesthetic evaluation of photos," *Neurocomputing*, vol. 172, pp. 244–252, 2016.

[5] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1 (CVPR'06)*, vol. 1, pp. 419–426, IEEE, New York, NY, USA, 2006.

[6] L.-K. Wong and K.-L. Low, "Saliency-enhanced image aesthetics class prediction," in *Proceedings of the 2009 16th IEEE International Conference on Image Processing ICIP 2009*, pp. 997–1000, Cairo, Egypt, November 2009.

[7] C. J. Cela-Conde, F. J. Ayala, E. Munar et al., "Sex-related similarities and differences in the neural correlates of beauty," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 106, no. 10, pp. 3847–3852, 2009.

[8] A. Forsythe, M. Nadal, N. Sheehy, C. J. Cela-Conde, and M. Sawey, "Predicting beauty: fractal dimension and visual complexity in art," *British Journal of Psychology*, vol. 102, no. 1, pp. 49–70, 2011.

[9] M. Nadal, E. Munar, G. Marty, and C. Cela-Conde, "Visual complexity and beauty appreciation: Explaining the divergence of results," *Empirical Studies of the Arts*, vol. 28, no. 2, pp. 173–191, 2010.

[10] R. Datta, J. Li, and J. Z. Wang, "Algorithmic inferencing of aesthetics and emotion in natural images: an exposition," in *Proceedings of the IEEE International Conference on Image Processing (ICIP '08)*, pp. 105–108, IEEE Press, San Diego, Calif, USA, October 2008.

[11] P. Machado, J. Romero, M. Nadal, A. Santos, J. Correia, and A. Carballal, "Computerized measures of visual complexity," *Acta Psychologica*, vol. 160, pp. 43–57, 2015.

[12] X. Tang, W. Luo, and X. Wang, "Content-based photo quality assessment," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1930–1943, 2013.

[13] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subjec," in *Computer Vision – ECCV 2008*, D. Forsyth, P. Torr, and A. Zisserman, Eds., pp. 386–399, Springer, Berlin, Germany, 2008.

[14] J. Neyman, "Basic ideas and some recent results of the theory of testing statistical hypotheses," *Journal of the Royal Statistical Society*, vol. 105, pp. 292–327, 1942.

[15] N. Orlov, L. Shamir, T. Macura, J. Johnston, D. M. Eckley, and I. G. Goldberg, "WND-CHARM: multi-purpose image classification using compound image transforms," *Pattern Recognition Letters*, vol. 29, no. 11, pp. 1684–1693, 2008.

[16] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 245–271, 1997.

[17] C. Fernandez-Lozano, J. A. Seoane, M. Gestal, T. R. Gaunt, J. Dorado, and C. Campbell, "Texture classification using feature selection and kernel-based techniques," *Soft Computing*, vol. 19, no. 9, pp. 2469–2480, 2015.

[18] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[19] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data," *Knowledge and Information Systems*, vol. 34, no. 3, pp. 483–519, 2013.

[20] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*, Springer-Verlag, New Jersy, NJ, USA, 2006.

[21] S. Maldonado, R. Weber, and J. Basak, "Simultaneous feature selection and classification using kernel-penalized support vector machines," *Information Sciences*, vol. 181, no. 1, pp. 115–128, 2011.

[22] R. R. Hocking, "The analysis and selection of variables in linear regression," *Biometrics*, vol. 32, no. 1, pp. 1–49, 1976.

[23] G. Tsiliki, C. R. Munteanu, J. A. Seoane, C. Fernandez-Lozano, H. Sarimveis, and E. L. Willighagen, "RRegrs: An R package for computer-aided model selection with multiple regression models," *Journal of Cheminformatics*, vol. 7, no. 1, pp. 1–16, 2015.

[24] S. Garcia, A. Fernandez, J. Luengo, and F. Herrera, "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: experimental analysis of power," *Information Sciences*, vol. 180, no. 10, pp. 2044–2064, 2010.

[25] C. Ambroise and G. J. McLachlan, "Selection bias in gene extraction on the basis of microarray gene-expression data," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 99, no. 10, pp. 6562–6566, 2002.

[26] K. Hechenbichler and K. Schliep, "Weighted k-nearest-neighbor techniques and ordinal classification," *Sonderforschungsbereich*, vol. 386, 2004, Discussion Paper 399.

[27] W. Liu and S. Chawla, "Class Confidence Weighted kNN Algorithms for Imbalanced Data Sets," *Advances in Knowledge Discovery and Data Mining*, vol. 6635, pp. 345–356, 2011.

[28] https://CRAN.R-project.org/package=gbm.

[29] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics and Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[30] https://projecteuclid.org/euclid.aos/1016218223.

[31] D. F. McCaffrey, B. A. Griffin, D. Almirall, M. E. Slaughter, R. Ramchand, and L. F. Burgette, "A tutorial on propensity score estimation for multiple treatments using generalized boosted models," *Statistics in Medicine*, vol. 32, no. 19, pp. 3388–3414, 2013.

[32] C. Fernandez-Lozano, M. Gestal, C. R. Munteanu, J. Dorado, and A. Pazos, "A methodology for the design of experiments in computational intelligence with multiple regression models," *PeerJ*, vol. 2016, no. 12, 2016.

[33] M. M. Marin and H. Leder, "Examining complexity across domains: relating subjective and objective measures of affective environmental scenes, paintings and music," *PLoS ONE*, vol. 8, no. 8, Article ID e72412, 2013.

Hindawi
Complexity
Volume 2019, Article ID 7293193, 21 pages
https://doi.org/10.1155/2019/7293193

Research Article

Image Evolution Using 2D Power Spectra

Michael Gircys and Brian J. Ross

Brock University, Department of Computer Science, 1812 Sir Isaac Brock Way, St. Catharines, ON, Canada L2S 3A1

Correspondence should be addressed to Brian J. Ross; bross@brocku.ca

Received 2 August 2018; Revised 23 November 2018; Accepted 6 December 2018; Published 2 January 2019

Academic Editor: Colin Johnson

Procedurally generated images and textures have been widely explored in evolutionary art. One active research direction in the field is the discovery of suitable heuristics for measuring perceived characteristics of evolved images. This is important in order to help influence the nature of evolved images and thereby evolve more meaningful and pleasing art. In this regard, particular challenges exist for quantifying aspects of style and shape. In an attempt to bridge the divide between computer vision and cognitive perception, we propose the use of measures related to image spatial frequencies. Based on existing research that uses power spectral density of spatial frequencies as an effective metric for image classification and retrieval, we posit that Fourier decomposition can be effective for guiding image evolution. We refine fitness measures based on Fourier analysis and spatial frequency and apply them within a genetic programming environment for image synthesis. We implement fitness strategies using 2D Fourier power spectra and phase, with the goal of evolving images that share spectral properties of supplied target images. Adaptations and extensions of the fitness strategies are considered for their utility in art systems. Experiments were conducted using a variety of greyscale and colour target images, spatial fitness criteria, and procedural texture languages. Results were promising, in that some target images were trivially evolved, while others were more challenging to characterize. We also observed that some evolved images which we found discordant and "uncomfortable" show a previously identified spectral phenomenon. Future research should further investigate this result, as it could extend the use of 2D power spectra in fitness evaluations to promote new aesthetic properties.

1. Introduction

1.1. Overview of Problem. Digital art brings to mind many wide and varying concepts and examples, with many digitally produced, original pieces finding their own acclaim [1, 2]. It is trivial for software to precisely replicate a digital image. On the other hand, we find it difficult to autonomously produce new images which share similar visual characteristics with images provided. Forming correct abstractions between digital data and their visual interpretations is an ongoing challenge covering many fields of study [3–6].

We focus on procedural textures, which are images generated with mathematical formulae and/or algorithms [7]. The terms "images" and "textures" are used interchangeably. Texture synthesis shows its use in applications ranging from interactive art systems [8], adaptive image filters [9], camouflage generation [10], and game asset generation [11] amongst others.

The ability to form minor alterations in these procedures allows us to easily make changes in a structured manner. However, it may not always be clear *a priori* how these changes will come to manifest. By combining together parts between the better performing generated images, we may gradually refine them and allow them to exceed the quality of any single prior image. With this process of evolutionary refinement, we are able to explore many similar images which can feature novel and creative variation. A technique to capture and replicate spatial properties would be of great benefit for improving these existing systems or expanding to new applications.

Evolutionary algorithms (EA)—and notably genetic programming (GP)—are able to nonexhaustively explore the space of possible images with little explicit understanding of how to affect high-level image changes [12–15]. Perhaps the most critical component in all EAs is the fitness measure, defining the metaheuristic which guides the search to optimal solutions. With image synthesis, a bridge is needed to cross the divide from computer vision, information theory, and computational intelligence attributes we can evaluate from our rendering, to the psychological and cognitive understandings of perception.

With evo-art, we are often attempting to recreate characteristics of a target image, and not to precisely duplicate it. The idea of evolving near-matches, or "variations on a theme", has been a goal in many previous applications [16–18]. Using an evolutionary approach, exact matches are possible for simple images, but become rather difficult for more complex targets.

In investigating the existing measures that can be computed from a rendered image, measures related to power spectral density appear to be promising. Estimates of power spectral density are based on the discrete Fourier transform of a signal, a measure of power across each component frequency. For 2D applications, a radial average of the 2D DFT coefficients with common polar distance (same spatial frequency) can be obtained for a more robust, abstract measure. A number of papers on image analysis/retrieval [4, 5, 19, 20] have been found which use this to more effectively classify images based on computationally tricky but perceptively obvious attributes (*i.e.,* Eastern versus Western art; Portrait versus Sketch versus Landscape). Despite this, little can be found relating to the use of power spectra for evolutionary art.

Power spectral density also plays a key role in spatial frequency theory. The theory purports that a human or animal visual cortex operates through coded signals in relation to observed spatial frequencies (in contrast to edge and line detection which can be prominently found in wavelets) [21–25]. An interesting adaptation of this research enables the identification of uncomfortable images through contrast and frequency analysis [3]. Power spectra of an image's luminance were investigated, and certain frequency octaves were found to provide higher ratings of perceptual discomfort. We find numerous motivations toward the exploration of power spectral density as an art fitness measure, and promise in modelling perceptual spatial characteristics.

1.2. Goals. With spatial frequency being one of the more human-intuitive measures for shape and composition, and with the amount of existing research linking the measure to human perception, this paper shows its potential as a tool for guiding evolutionary textures. Our goal is to explore the use of these measures in evolutionary texture synthesis and evaluate their utility in production of digital evolutionary art. We consider our models of shape from a target image for use as a guide when evolving new images. It is hoped that by capturing and reproducing key spatial attributes of the image, we can see novel images with similar properties emerge in a creative exploration.

Our research presents a pair of milestones. Using genetic programming, we produce grayscale textures and explore the ability of Fourier-based fitness measures to replicate spatial properties of target images. The focus on grayscale images simplifies the texture formulae evolved, and permits experiments to concentrate on shape information. We then explore the use of these measures for colour image synthesis. Most evo-art systems use colour, and so it is important to examine the applicability of our Fourier analyses to the colour domain. Doing so helps establish the utility of Fourier shape analysis as a tool for serious applications in evolutionary art.

1.3. Organization of Paper. The paper is organized as follows. Section 2 reviews the Fourier transform and its application toward 2D images. Section 3 discusses some of the important research literature of relevance to this paper, with a focus on evolutionary textures, and application of power spectral density measures. We outline the details of our experimental system in Section 4, and summarize the key findings of our initial experiments in Section 5. Later work with adaptations toward evolutionary art is discussed in Section 6. Conclusions are given in Section 7.

The paper presumes familiarity with genetic programming [14]. Further details of this research are in [26].

2. Background

2.1. Fourier Transform. The following briefly outlines some of the main technical details of Fourier analyses. A complete introduction is beyond the scope of this paper. We refer the reader to detailed discussions in [27–29].

Fourier analysis is a well-known tool which sees substantial use in signal processing applications [28]. The Fourier transform converts a signal with samples based on amplitude at points in time, to a representation which shows the power and phase of the signal's constituent frequencies. The Fourier transform translates a signal into a sum of sinusoids, where the frequency of each periodic term relates to a component frequency found in the signal. The result of such a decomposition is typically encoded as a complex number for each frequency (see (1) to (3)).

$$f(t) = \sum_{n=-\infty}^{\infty} C_n e^{in\omega t} \tag{1}$$

$$C_n = \frac{1}{T}\int_0^T f(t)\, e^{-in\omega t} dt \tag{2}$$

$$= \frac{1}{2}(a_n - ib_n) \tag{3}$$

The real part of the coefficient (a_n) scales each term and may maintain its definition as the amplitude of the particular frequency. The additional imaginary component of the coefficient (b_n) can be used in conjunction with the real component to recover the phase of the frequency, as declared through the complex phase angle.

Adapting the Fourier transform to a 2D image can be done by applying the discrete Fourier transform (DFT) on each index of the first dimension, and then again along each row of the results. This gives us the amplitude and phase of how each frequency contributes to the total 2D signal. In applications with images, we often see most of the high-energy coefficients appear around the central positions and main axes of the shifted DFT [29], as seen in Figure 1.

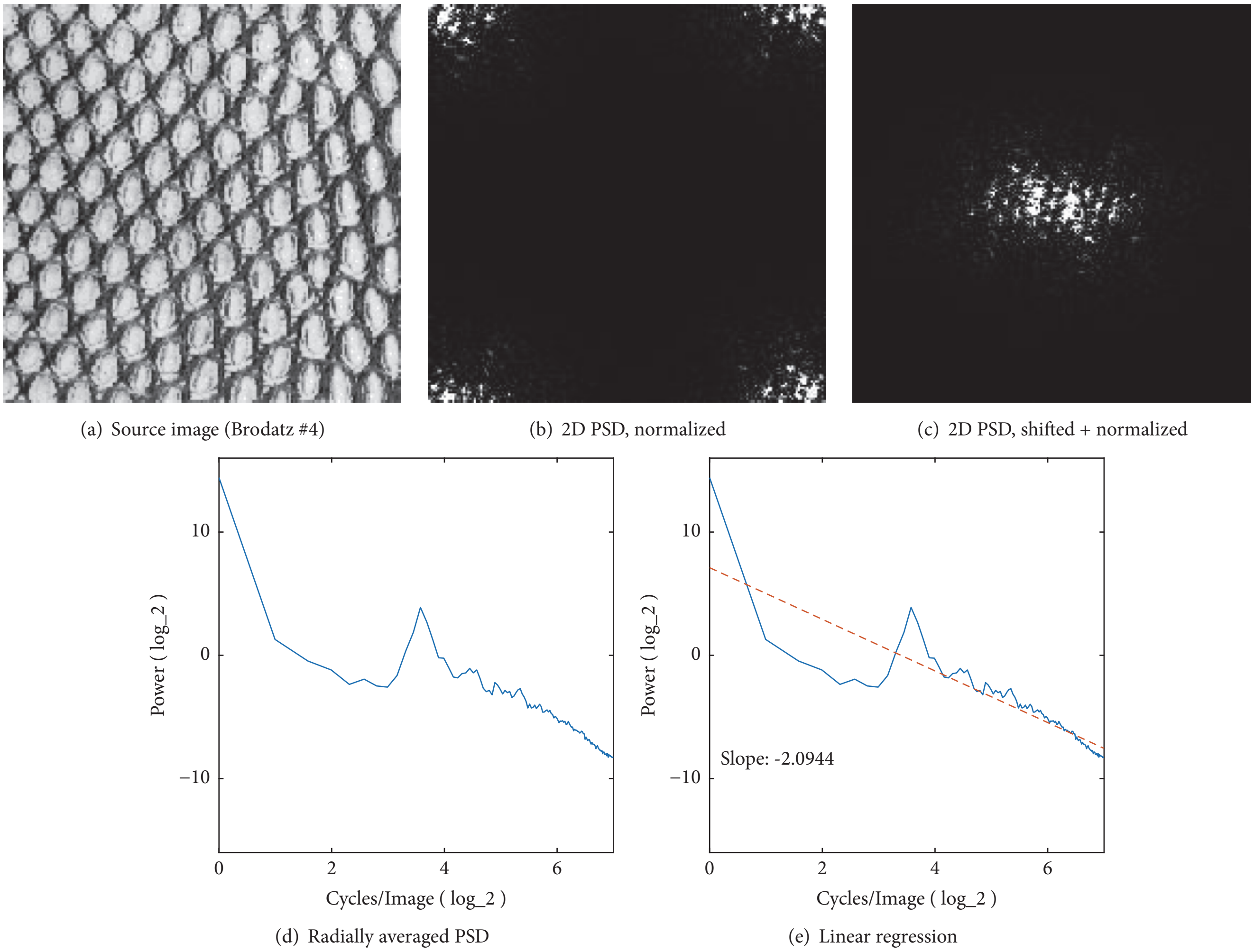


Figure 1: Power spectra pipeline. The source image in subfigure (a), undergoing 2D Fourier analysis, and its power spectral estimation shown in subfigure (b). As it provides for more interpretable charting and easier radial estimation, we "center" the coefficients by shifting them to diagonally opposite quadrants as seen in subfigure (c). We can then reduce dimensionality and produce useful aggregates by using radial averaging measures (d) and subsequent regressions (e).

Where the amplitude of an audio signal may have an intuitive correspondence with sound wave pressure, amplitudes for a 2D image will be a measured in relation to their pixel intensity, or as is typically the case in colour images, the intensity across a particular colour channel.

While the Fourier transform can scale to higher dimensional signals, the use of DFT for colour textures is still potentially problematic [30]. In consideration of applying the DFT to colour channels in isolation, we should note that spatial properties are not necessarily clear from average intensity nor from inspection of individual colour channels. The related quaternion Fourier transform [31] might assist in this matter.

2.1.1. Power Spectral Density. The power spectral density (PSD), or power spectrum, is a measure of the power across the frequency domain of a signal. We can acquire an estimate of the PSD P_j at frequency j, by multiplying the Fourier terms C_j by their complex conjugate $\overline{C_j}$ and scaling by the number of samples n to produce a periodogram [32]. Due to the simple, real-valued coefficients of our image signal, we can simplify this to normalizing and squaring the real part of the DFT, as in (4).

$$P_j = \left(\frac{C_j\overline{C_j}}{n^2}\right) \tag{4}$$

$$= \left(\frac{\left|C_j\right|}{n}\right)^2 \tag{5}$$

For a 2D signal, we will be interested in the radial average of this measure, requiring us to shift the quadrants of our estimate, and then interpreting the average in a polar coordinate system. An overview of the steps in our measurement pipeline is shown in Figure 1. Between the DFT and the radial averaging methods, the power spectral estimate measure has the benefit of being approximately equal across rotation, and

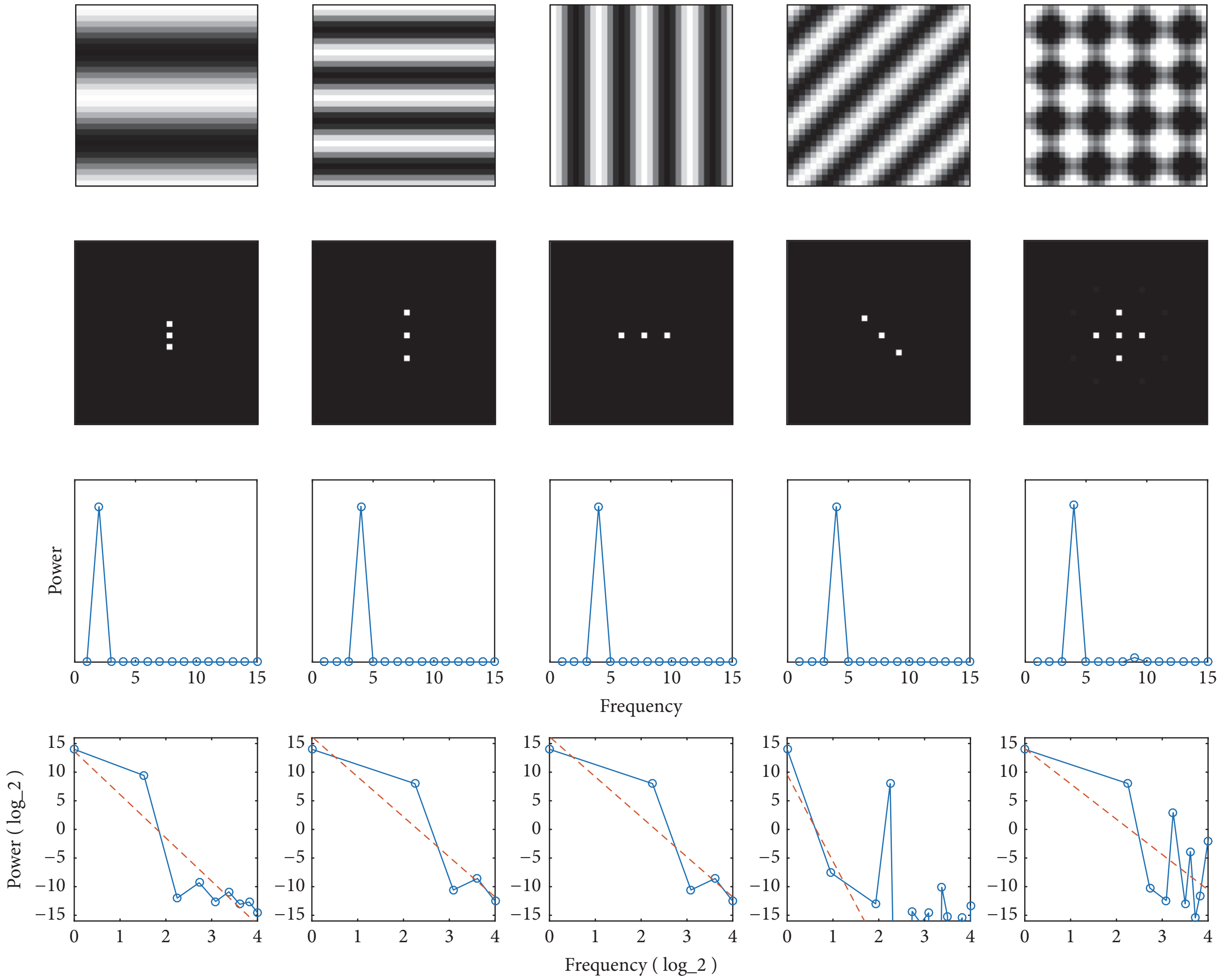


Figure 2: Power spectra interpretation and reconstruction. Rows from top to bottom: source image, shifted, and normalized constituent FFT power coefficients, radially averaged power spectra, and radially averaged power spectra plotted in a log-log scale ($\log_2$) with linear regression.

preserving shape across resolution. This measure relates to the contrast of luminance intensity, and we may also see a relation with image complexity. A further abstraction is to take a linear regression of the averaged power spectral density. While a display of the 2D power coefficients may more accurately represent the true power spectral density of a 2D signal, we find in some of the literature (*i.e.,* [4, 20]) that "power spectral density" and related terms often refer to the radial average or similar abstractions.

Figure 2 illustrates various representations of an image with a single component frequency. Shifting from the first to second column of the figure, we can see that lower frequency (those which have larger periods/cycles over greater areas of the image) is contained at the center of the shifted FFT power coefficient display. The first column shows a wave whose period is half of the canvas (input signal), and so the charted radially averaged power spectrum shows high power at a frequency of 2. As we move to the outer edge of the power coefficient display, we find the powers of increasing frequency ranges being displayed. The fifth column faintly shows a suitable example of minor aliasing artefacts having both lower power and higher frequency as we move from the key frequencies toward the image edges. We can also observe that the orientation of the wave-like pattern in the top image corresponds to the angle (from center) of the coefficient responsible for the effect, while still maintaining a distance (from center) corresponding to the actual frequency. Observing the subsequently charted radially averaged power spectra plots, we can see that all have a high power at frequency 4. Finally, we can see the multiplicative combination of the two component frequencies in the last column, as a grid begins to form with both horizontal and vertical frequency, again reflected in the power coefficient display. The final row of the figure displays the radially averaged power spectrum in a log-log scale, to assist in showing the much larger 0^{th} coefficient, and the more subtle changes in the lower-powered high frequencies. However, in simple images, there may not always be power at every frequency. A problematic consequence of this is that these frequencies cannot be charted in a log-log scale, and may affect the results of any regression, as is visible in the figure.

3. Literature Review

Although evolutionary algorithms have been applied to many forms of art over the years, we focus on literature involving the targeted evolution of procedural textures.

3.1. Spatial Measures. The need for measures permitting comparison of spatial properties tends to get resolved through one of two main concepts. Common approaches either extracted key features (and their positions) from a source or target image, or performed some type of frequency analysis. Many early attempts to capture spatial aspects for image database systems relied on basic algebraic and statistical measurements across intensity. The QBIC Project (which explored image querying through use of colour, texture, and shape measures) proposed spatial measures derived from capturing intensity areas, circularity, eccentricity, axis orientation, and algebraic central moment information [6].

A notable paper pertaining to image retrieval was published by Jacobs *et al.* [33], in which the proposed algorithm was capable of efficiently extracting the key coefficients from wavelet analysis. Extracted coefficients are limited to the K greatest absolute values, before being quantized and compared for mismatch. While the algorithm may have been intended for a retrieval system, the comparative abilities of the measure proved effective in guiding evolutionary systems. In [33], the set of coefficients were "truncated" by zeroing all but the top K greatest absolute value coefficients. Following this was a "quantization", setting all nonzero components into their sign of $\{-1, +1\}$. The total error between images could then be found by summing of differences between each truncated, quantized coefficient position. This quantization scheme was found to be quite beneficial; despite the resulting loss of precision, as "the mere presence or absence of such features appears to have more discriminatory power for image querying than the features' precise magnitudes" [33].

3.2. Evolutionary Textures. The use of evolutionary algorithms for texture synthesis was pioneered by Sims [13], and used interactive user guidance, which enabled a user to gradually manipulate sets of graphical shaders to produce images fitting a desired aesthetic.

An early attempt in the transition to unsupervised approaches came from Baluja *et al.* [15]. Simple topologies of artificial neural networks were used in an attempt to learn a user's aesthetic preferences by training against user ratings and groups of raw pixel values. This approach saw some shortcomings, but highlighted the need for abstracted image measures to be used as guides. The idea of learning aesthetic preferences through neural networks has since been revisited with the inclusion of multiple abstracted image measures with some reported success [34].

A critical successor to Sims' work was the Genshade system by Ibrahim [16]. Genshade introduced unsupervised, automatic fitness evaluation of images as generated by evolved Renderman shaders. Various image analyses were compared between the evolved images and a provided target image. These measures were used in lieu of user input to guide the evolution of textures toward those showing similar visual characteristics of the targeted image.

The Gentropy system by Wiens and Ross [17] expanded upon the unsupervised approach of Genshade by providing additional image analysis measures, and use of a simple procedural texture language, in contrast to Genshade's evolution of high-level Renderman shaders. A suite of image analyses were performed during fitness evaluation, which benefited with the use of island-model parallelism for maintaining diversity and accelerating the quality of evolved results. Gentropy was later enhanced in [35] by replacing island-model evolution with multiobjective evaluation, by treating the different image analysis tests as separate objectives for Pareto ranking.

Genshade [16] and Gentropy [17] employ the techniques from [33], where spatial features were compared via these extracted coefficients from wavelet measures. The technique appears to have been successfully adapted for use with texture synthesis. Results of wavelet analyses in both systems were positive, although a comprehensive investigation regarding the extent of their abilities was not undertaken.

More recently, there have been developments in using aesthetic modelling to guide image evolution [36–39]. Aesthetic modelling is a pioneering frontier for art and image analyses, and proposed models are not yet mature enough to be comprehensive theories of artistic beauty and aesthetics. Nevertheless, these efforts attempt to use higher-level image analyses as guides for evolution, which contrasts to the lower-level image processing used by systems like Genshade and Gentropy.

Recent work by Tanjil [40] uses ideas from deep learning to guide evolutionary image synthesis. A heuristic is proposed that enables activation nodes of a deep convolution neural network (trained for classification) to be identified for use by fitness evaluation. Using a set of images sharing desired visual features, the heuristic determines the activation nodes of the network most likely to be activated by the visual characteristics of interest. These nodes are then used as guides by fitness. A number of experiments showed that the genetic programming system was able to evolve images which shared desired properties of target images, such as shape and colour. Tanjil concludes that, as deep learning networks become better understood, they may be even more effectively exploited by evo-art systems.

While these and other systems attempt to capture spatial attributes, that was only a part of their purpose as a more generic art system. There was no extensive evaluation of their spatial guidance capabilities, and the use of Fourier analysis in texture synthesis or aesthetic modelling has been left largely unexplored.

Further examples and surveys of evolutionary art can be found in [1, 2], and contemporary research is published at the annual EvoMusArt conference (http://www.evostar.org/).

3.3. Limitations. Although the use of wavelet-based analysis showed effectiveness, alternative approaches are possible. One considered problem with a frequency analysis approach was in the inability to effectively handle images with multiple colour channels [30]. One potential solution to this was proposed through the use of the quaternion Fourier transform [31], which does not have a direct equivalence with wavelet analysis.

A criticism common to all types of frequency analysis remains in the fact that a perfect solution would exactly replicate the target image [41]. In evolutionary art, we never desire to make a reproduction of a given image. Rather, we only want to capture key characteristics of an image, and explore the landscape of possible solutions which are in some way similar. While fitness evaluations could be adjusted to prefer some amount of error, we found that there is often still sufficient challenge presented to our system outside of toy problems, permitting for novel solutions to emerge while we pursue higher numerical accuracy.

4. System Design

There are two key components which form the core of our experimental system. The first component is a library which could process an image to provide the power spectral density (PSD), regression, and other FFT related measures. The second, and largest, component is the evolutionary system which used genetic programming to evolve and synthesize procedural textures.

4.1. Power Spectral Density Measures. A number of PSD-related calculations were required for this research. For example, the 2D power coefficient matrices, the radially averaged power spectral density, and its linear regressions. MATLAB [42] (release 2016a) was used to assist with computation of power spectral density measures. MATLAB allowed us to generate native C code, which was integrated into the Java-based evolutionary system (Section 4.2) through use of the Java Native Interface (or commonly, JNI) framework.

For the experiments using regression measures of the radially averaged power spectral density, the regression was obtained first by converting the power measures to a log-log scaling, to better match the conventional practices seen in the literature. Charting of PSD throughout this paper uses $\log_{10}$ scaling to remain consistent with other charted scales, though evaluations used for the various applicable experiments have used a $\log_e$ scaling. For a linear regression, the slope measures should remain identical across log bases, though the offset will vary. Regressions were found by using MATLAB's `polyfit` function, which itself performs a least-squares error fit. While uncommon for natural images, some abstract images produced by our system were found to have no power at certain frequencies. To lessen the biased effects of these values from the regression, any infinite or invalid power measures were removed from the set of points considered during the regression.

We decided to forgo any image windowing functions prior to sending the image data through the DFT and PSD measure pipelines. The use of a windowing function has been advised for nonregular signals, such as typical nonrepeating images, to reduce heavy artefacts in the decomposition. Specific window functions and parameters would be dependent on the expected signal. However, initial trials using windowing did not significantly impact our results, and so windowing was henceforth ignored.

In summary, tests found that our library produced results closely matching existing literature, and specifically those from Graham *et al.* [20].

4.2. Genetic Programming Engine. The evolutionary art system we used to generate textures is a custom extension of the ECJ system (version 23), a Java-based system for genetic programming and other techniques [43].

We used a genetic programming tree representation to evolve symbolic expressions for procedural textures. Much of our early experimentation focused on spatial attributes of an image. We found that grayscale textures were not only adequate, but were indeed preferable over the artistic colour texture renderings. To suitably represent this, GP individuals needed only a single tree to evaluate luminosity or intensity. Later experimentation expanded to colour textures, and we consequently expanded our individuals to hold 3 trees; one tree was used for each colour channel in the RGB colourspace.

The wall-clock run times for the system configured for basic grayscale textures were found to be approximately 45 minutes per run, when executed using a single thread of an AMD FX-8350 processor. In this configuration, multiple runs were evaluated concurrently. With the parallel nature of the system, we could see substantial reductions in single-run execution time if reconfigured to use multiple threads. The introduction of noise operators and RGB colour channels each increased runtime by factors of approximately 6 and 3, respectively. Coloured textures using noise language operators required an approximate average of 12 hours for completion of a run.

4.2.1. GP Parameters. Table 1 lists the GP parameters normally used in our experiments. Although most are standard in the literature [14], a few require explanation. Three variants of ephemeral random constants (ERCs) were included corresponding to orders of magnitude, and each of the ERC nodes are instantiated to random values within their respective ranges. The introduction of ephemeral value mutation allowed for the randomized constants to be slightly altered by 1%, which permitted for finer adjustments to the rendered image. The ERC mutation operator had been included at a probability of 10% and was responsible for a proportional decrease in likelihood to execute the crossover operator. So as to remove the possibility of losing the best found individual in a generation, we allowed elitism for the single best individual of a generation to be retained unaltered in the subsequent generation.

The termination criteria for a run were the completion of 100 generations. While "perfect" individuals had been produced for some simple compositional targets, this was

Table 1: Genetic programming engine parameters overview.

Parameter	Value
Runs	30
Generations	100
Population Size	1000
Elitism	1
Sum-of-Ranks Fitness	
Diversity Penalty, Initial	10
Diversity Penalty, Increment	10
Generation 0	
Builder	Ramped Half & Half (see [14])
New Node Depth	2 . . . 6
Grow Probability	50%
Reproductive Operators	
Crossover	70%
Mutation	20%
ERC Mutation	10%
Crossover Max Depth	17
Mutation Max Depth	17
Mutation New Node Depth	5
Selection Method	Tournament
Tournament Size	3

otherwise a difficult problem, where finding such a “perfect” solution was not typically expected.

4.2.2. Texture Languages. The GP language is in Table 2. Standard mathematical operators were used, as well as specialized texture generating primitives. Optimized Perlin and simplex noise generators have been borrowed from [44, 45] respectively. The fractalsum, turbulence, and marble noises have been based on the Perlin noise implementation as originally conceived. For these noise variants, coordinate scaling had been used to ensure noise is applied across the $[-1, 1]$ rendering window. Initial experiments in Section 5 excluded the spatial and noise operators.

4.3. Multiobjective Evaluation. Some problems permit us to evaluate solutions with a single measurement, for example, the overall error in a regression problem. However, there are problems where multiple criteria are necessary. These metrics can be independent, or can interact in complex, nonlinear ways. Reconciling such factors into a single metric score, for example, by a weighted sum, can be challenging to do effectively, and detrimental to search. The field of multiobjective optimization is concerned with problems such as these, in which multiple objectives are involved in defining the search criteria for a problem [46].

A popular scheme for scoring multiobjective problem spaces is Pareto ranking [47]. With Pareto, individuals are scored in relation to the others in the population. Unfortunately, Pareto ranking is not suitable for problems involving more than 3 objectives.

Our system uses the sum of ranks (or average rank) strategy, which was devised for multiobjective problems involving a high number of objectives (termed “many-objective” problems) [48, 49]. Sum of ranks encourages solutions to perform well across all considered objectives. It is also effective for problems having a large number of objectives (unlike Pareto ranking). The sum of ranks approach has been found effective in evolutionary art applications [35, 39].

Table 3 illustrates the calculations for sum of ranks. After obtaining the raw measures (O_i) for each fitness objective, each measurement is separately ranked (R_i) relative to other individuals in the population. The rank scores are normalized (N_i) by dividing each R_i by the maximum rank value for that objective. The normalized ranks are summed for each individual, resulting in a fitness measure. The sum of ranks score denotes an individual’s relative performance of its objectives relative to the population at large. The final column Rank shows the relative fitness quality of each individual in the population. For example, individual #1 has the best score in each objective relative to the rest of the population, and thus has the best (lowest) sum of ranks. Individual #3 has an extremely poor score of 99 for objective 2. However, this raw score is converted to a rank of 5, and therefore does not unduly penalize the final ranking.

By using sum of ranks in our system, we are able to maintain a consistent diversity penalty scheme across all experiments. For individuals whose ranks in all objectives are identical, the second individual would have a penalty of 10 added to each of their ranks. Additional individuals found with the same scores as the first would incrementally receive an additional penalty of 10 rank points (the fourth common individual would receive a total of +30, and so on). These penalties are used to maintain genetic diversity in the population by penalizing identical results.

5. 2D Fourier Fitness Strategies

5.1. Simple Regression and Error. We first considered the error between FFT decomposition from evolved individuals and its target at a high level of abstraction. Beginning with the technique common in the literature (e.g., [4]), we considered a fitness scheme which measured the difference between slopes found through linear regression.

Measures of linearly regressed, radially averaged power spectra displayed some effectiveness previously with classification and retrieval. Consequently, evaluating fitness through this measure seemed like a promising start. Previous literature showed an improved ability to distinguish genre by incorporating this measure, and it was hoped that some spatial property capable of distinguishing these genres might emerge in our evolutionary synthesis.

In selection of a target set (Figure 3), we focused our efforts on aspects of spatial composition similarity. Though visually simple, the target images included basic compositions which might be used for evolutionary art.

Some concerns arose early into the process of constructing the linear regression module for our GP system. While much of the earlier explored work focused on evaluating natural images or complex art pieces, little investigation had been done into simple synthesized textures. In the process of

Table 2: Genetic programming engine base language overview.

Category	Arity	Display	Description
Variables	0	X, Y	Texel rendering coordinates ($-1 \leq X, Y \leq 1$)
		Rho	Polar coordinate; distance from $\{0, 0\}$
		Phi	Polar coordinate; angle about $\{0, 0\}$ to X axis
Ephemerals	0	E[1]	Ephemeral in range $[0, 1]$
		E[10]	Ephemeral in range $[0, 10]$
		E[100]	Ephemeral in range $[0, 100]$
Math	1	-	Negation / sign change
		abs	Absolute value / magnitude
		floor	Floor; lesser or equal whole integer
		ceil	Ceiling; greater or equal whole integer
		sin, cos, tan	Periodic, trigonometric functions
		sqrt	Square root
		exp	*e* (Euler's number) raised to the operand
		pow2, pow3	The operand raised to a fixed power of 2 or 3
		log_E, log_10	Natural log, and log of base 10
	2	+, -, *	Addition, subtraction, multiplication
		/	Safe division; a zero divisor returns zero
		max, min, avg	The greater, lesser, or mean of two operands
		pow	arg[0] raised to arg[1]
	3	lerp	Linear interpolation between arg[0] and arg[1] based on normalized (clamped to $[0, 1]$) arg[2]
Conditionals	4	IfGT	If arg[0] > arg[1] then arg[2], else arg[3]
Spatial	1	Circle	Gives 1.0 where *Rho* <= arg[0], otherwise 0.0
	3	Shift	arg[0] evaluated in rendering position shifted by arg[1] horizontally, and arg[2] vertically
		Tile	arg[0] evaluated in rendering position scaled and offset for a arg[1] × arg[2] window tiling
Noise	0	Simplex [†]	Simplex noise generator
		Marble [†]	Marble noise (see [7])
	1	FractalSum [†]	FractalSum/Smooth noise
		Turbulence [†]	Turbulence noise

[†] All noise functions include a variant symmetric about the X and Y axis. These variants would have a Sym prefix and function otherwise identical to the base function.

Table 3: Example of sum of ranks for a 3-objective problem. Lower objective scores and ranks are preferred. The maximum rank for each objective R_i used for normalization is in boldface.

	Objectives			Rank			Normalized Rank				Final
#	O_1	O_2	O_3	R_1	R_2	R_3	N_1	N_2	N_3	ΣN_i	Rank
1	1	1	3	1	1	1	0.25	0.2	0.33	0.78	1
2	2	2	4	2	2	2	0.5	0.4	0.67	1.57	2
3	2	99	3	2	**5**	1	0.5	1.0	0.33	1.83	3
4	4	4	4	3	3	2	0.75	0.6	0.67	2.02	4
5	6	7	5	**4**	4	**3**	1.0	0.8	1.0	2.8	5

charting the linearly averaged power spectra, and producing its regression, a transform into the log-log scale was required. Often, simple geometric images would result in frequencies with zero power. These anomalous frequencies needed to be removed, which could have an impact to the quality of regression.

Some example solutions for the slope results are shown in Figure 4. One positive aspect is that GP easily evolved images with a high degree of fitness to the targeted slopes. The slope measure alone was insufficient in capturing any sufficient amount of spatial details. We found our GP system invariably converged to visually simple textures. The fitness criteria was too easily satisfied, and language biases were prevalent through our choice of simple mathematical operators. Unlike the use of regressed slope in image classification where it was applied to highly defined image sets (artwork, natural photographs, *etc.*), GP was capable of finding trivial solutions with the given slope criteria. Other experiments

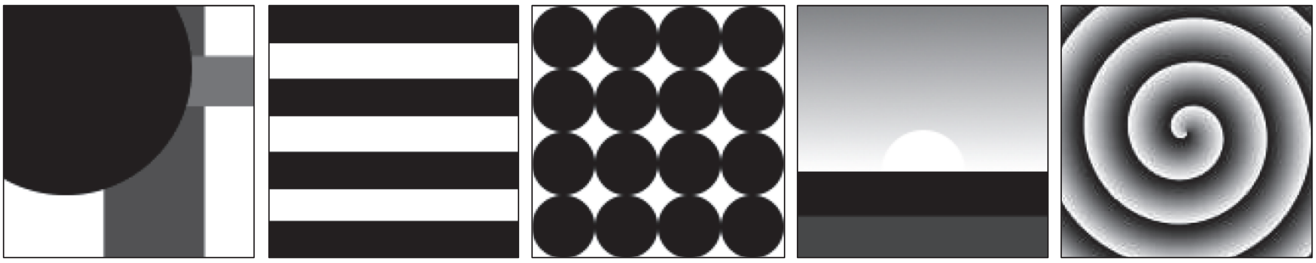
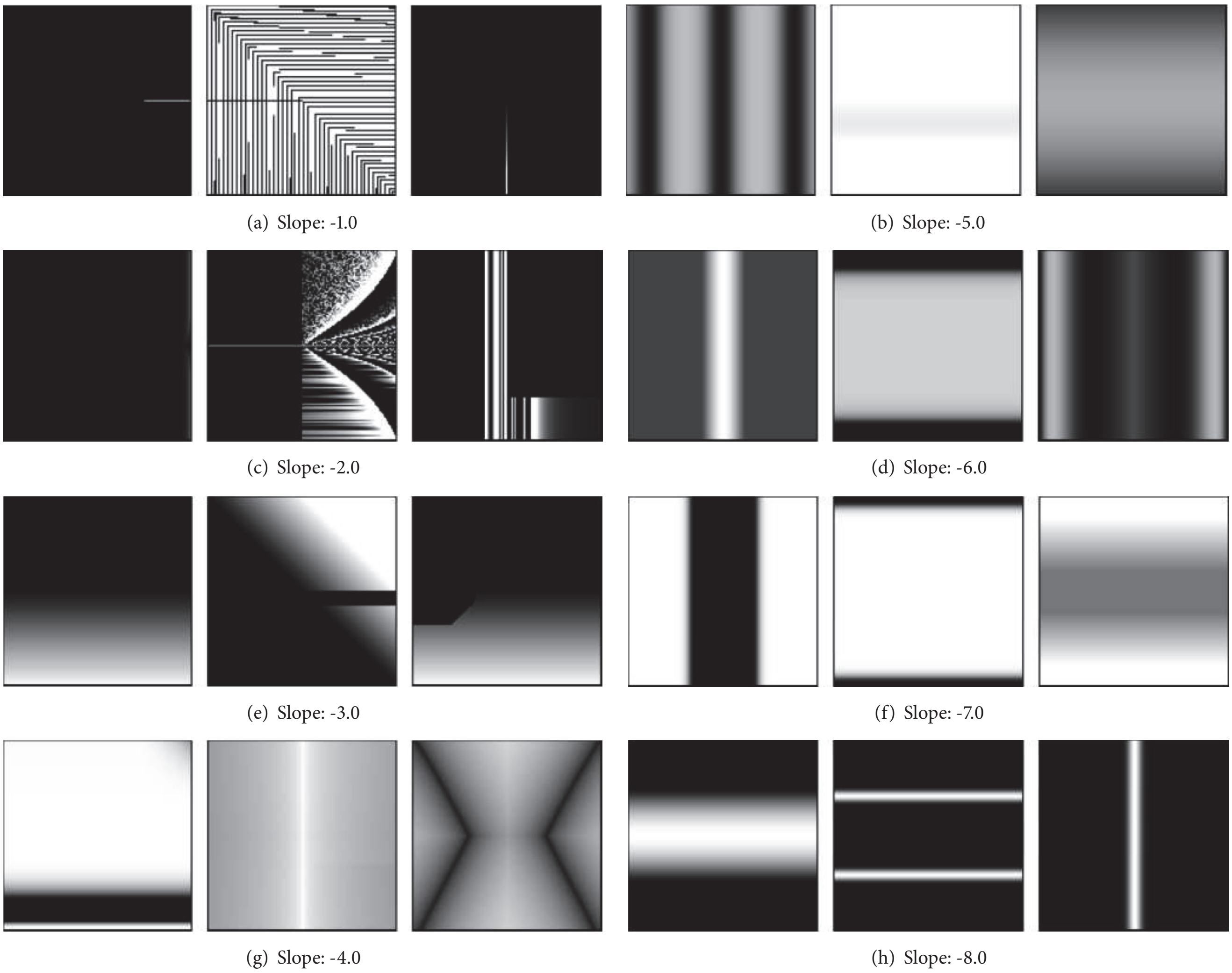


Figure 3: Compositional target set.



(a) Slope: -1.0

(b) Slope: -5.0

(c) Slope: -2.0

(d) Slope: -6.0

(e) Slope: -3.0

(f) Slope: -7.0

(g) Slope: -4.0

(h) Slope: -8.0

Figure 4: Regressed slope example evolved images. Target slopes were specified at regular integer intervals from −1.0 to −8.0. Best candidates per run had mean $\varepsilon < 1.0E^{-5}$, except for the last target, where it was found that errors greatly increased when target slope exceeded -7.0. These initial runs relied on Cartesian coordinate variables (omitting the polar coordinate variables).

using power spectra regressions and similar basic measures were performed with only modest improvements to results (see [26]).

5.2. Filtering Relevant Coefficients. A promising strategy for coefficient isolation in frequency analysis was found by Jacobs *et al.* [33] using wavelets (See Section 3.1). There were a few considerations to note before attempting similar schemes using Fourier transforms. A quantization to $\{-1, 0, +1\}$ was not as meaningful in the context of a Fourier transform, where power coefficients were strictly positive. Amplitude coefficients may have held negative values, but these could change sign when set with appropriate phase. We could truncate coefficients as per the paper, but the solution we

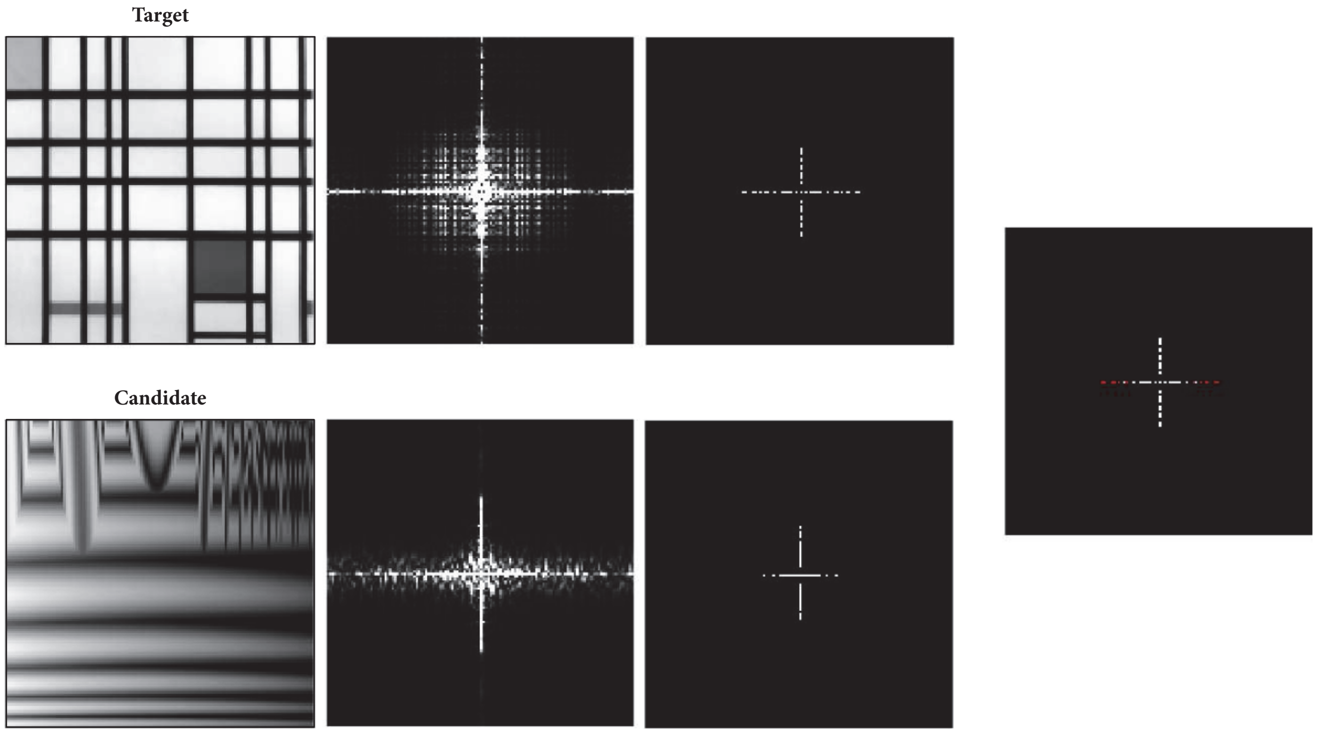


Figure 5: Truncation and quantization example. Target and candidate coefficient sets were reduced to the top $K = 50$ most powerful positions and then quantized to 0, 1, before checking for matches. The rightmost image shows the top K positions matching between target and candidate, with those absent from the target being displayed in red.

attempted instead quantized all remaining values in Boolean to 1. This effectively turned the score into a count of how many positions shared a top K coefficient between target and candidate. For a target and candidate of equal size, we ranked the target's coefficient positions by their power, and truncated all but the top K. Each candidate could then undergo the same coefficient ranking process, and check its top K for a nonzero value in the corresponding location of the target's truncated coefficients (see Figure 5).

While a wavelet decomposition would require further choices for wavelet type, decomposition type, and basis normalization schemes, Fourier compositions are constrained but simplified. A choice of K value was still required to determine the size of our coefficient truncation. Jacobs *et al.* found values of 40 to 60 performed well with their image retrieval data sets [33]. In our selection of a suitable K value, we considered possible reconstructions of the target images where power was removed from all but the top K positions. Prominent recreations began to form in the range of $K = [50, 150]$, where certain targets performed well with as low as $K = 10$.

5.3. Phase Refinement. A critical difference between the wavelet strategy of Jacobs *et al.* [33] using wavelets, and our adaptation with Fourier transforms, was the inherent removal of any spatial localization in our frequency analysis. When measuring coefficients, the index and position (the radial angle of the coefficient from center) encouraged evolution of component frequencies with similar placement. However, this tended to overlook how these component frequencies should be offset and overlap. The other key aspect of a Fourier transform, the phase component, must therefore be considered. By reincorporating phase into our fitness scheme, we provided further constraints on the location of where the component frequencies crest. See Figure 6 for examples showing the effect of phase in Fourier reconstruction.

We adapted the Jacobs *et al.* approach—or, top K mismatch— and considered the difference of phase angle for those top K positions. Being mindful that phase error should wrap about 2π, the maximum difference in phase angle should be π. We normalized the phase error to $[0, 1]$ and squared it for each of the top positions. This error was then used to slightly penalize the top matching positions if they are out of phase.

We separated the phase error component to its own sum of ranks fitness objective, and applied a scaling factor on the phase to prioritize the more visually prominent (powerful) components. This is more formally defined in (6) and (7) and was also used for the next experiment.

$$Error_{power} = \sum_{i=1}^{K} \begin{cases} 0.0, & T_i \in C \\ 1.0, & T_i \notin C \end{cases} \tag{6}$$

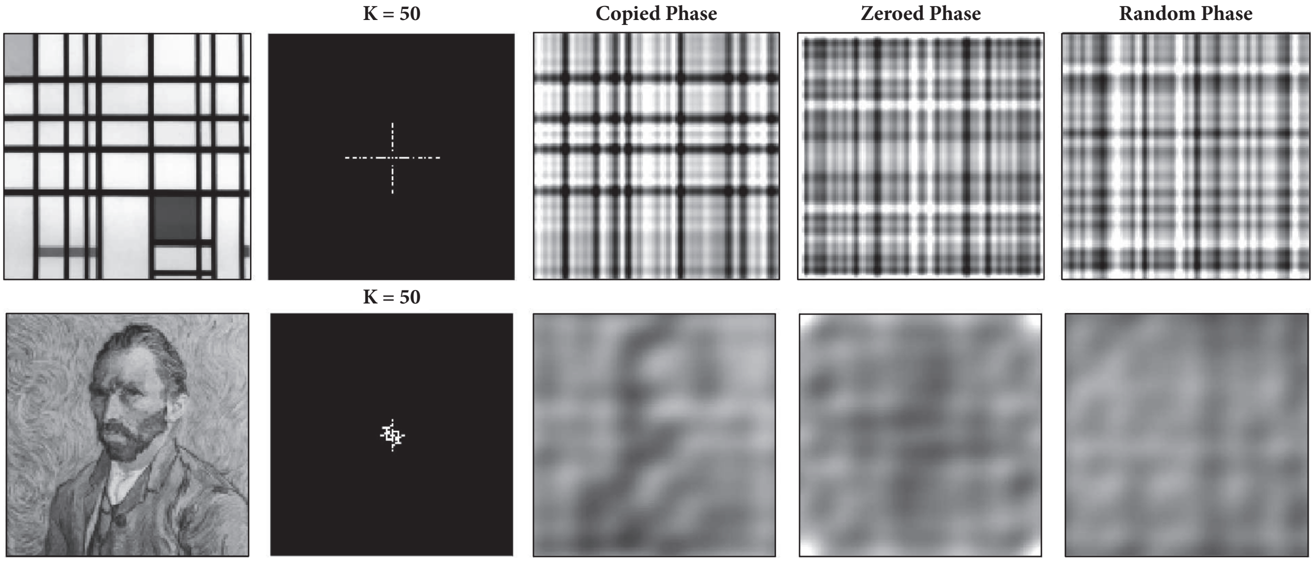


Figure 6: Reconstructing target images with varied phase. A pair of targets were chosen for reconstruction with a truncated set of Fourier amplitudes paired to different phase angle values. By zeroing power in all positions except the top $K = 50$ most powerful positions, we can see the most salient positions alongside the target in the second column. The third column is the inversed FFT reconstruction with power limited to these truncated positions. We also show reconstruction variants using the same truncated amplitude set, but with zeroed phase angles, or phase angles which have been produced randomly. This may adjust expectations for the types of images evolved when phase is not considered.

$$Error_{phase} = \sum_{i=1}^{K} \begin{Bmatrix} \left[\pi^{-1}\Delta\left(\theta\left(T,T_i\right),\ \theta\left(C,T_i\right)\right)\right]^2, & T_i \in C \\ 1.0, & T_i \notin C \end{Bmatrix} \cdot \left(\frac{K-i+1}{K}\right) \qquad (7)$$

The equation assumes an $n \times n$ power coefficient set, where T and C are the truncated set of coefficient positions for the target and candidate as ordered by power. We have $\Phi(V,p)$ and $\theta(V,p)$ return the power and phase angle respectively of the coefficients (complex/vectors) in set V corresponding to coordinate p. With a slight abuse in notation, we denote the coordinates of the i^{th} ranked position (by power) of a coefficient set as S_i.

We show our key results in Figure 7. Using our measure, we were able to evolve images which show variations of their targets' key features. Similar regions of intensity can be seen for Composition_01, consistent horizontal stripes are produced for Composition_06, and vertical regions and gradients can be found in Composition_09 (some of which capture the finer details near its center). To have the regions of intensity seen in compositions 01 and 09 reproduced, proper capturing and recreation of phase information would be required. The low phase error seen for these targets (Table 4) is reflected in their visual similarity. The curves produced for the spiral target of Composition_10 are also quite interesting; the target was expected to be more difficult to satisfy, be we find variations of the key radial aspects are reliably recreated despite slightly elevated fitness error. Some notable examples produced have been highlighted in Figure 8.

Composition_06 (horizontal stripes) evolved candidates which scored well with our measure, and certainly captured the idea of horizontal stripes, but were not as uniform as seen elsewhere (see [26]). Despite closely matching the top $K = 10$ coefficients with its target, many evolved candidates also held large amounts of power in other coefficients. We found this was mitigated by adjusting K (at the cost of increasing outlier results), or trivialized by reducing the GP language. Particular difficulty was seen with Composition_07 (circle grid), but for different reasons. With this target, the produced solutions had high levels of error through our fitness measure. Our GP system allowed for the easy formation of unit circles and lines along the dimension axes, which makes for an underwhelming capture of the grid and circular aspects desired.

Extended runs terminating at 200 generations were attempted with little change to image quality. We can find further improvements on the targets with circular composition aspects by adjusting our GP language (Section 6.1.1). While certain targets may have performed better individually with various adjustments to the fitness measure (see [26]), the results from the above measure (shown in Figure 7) performed generally well across the majority of our target images.

6. More Advanced Artistic Explorations

Whereas Section 5 considered greyscale image synthesis, this section expands the scope of image evolution by considering more complex colour images. We first consider enhancements and extensions to our GP language which may better reflect some of the more full-featured languages used for

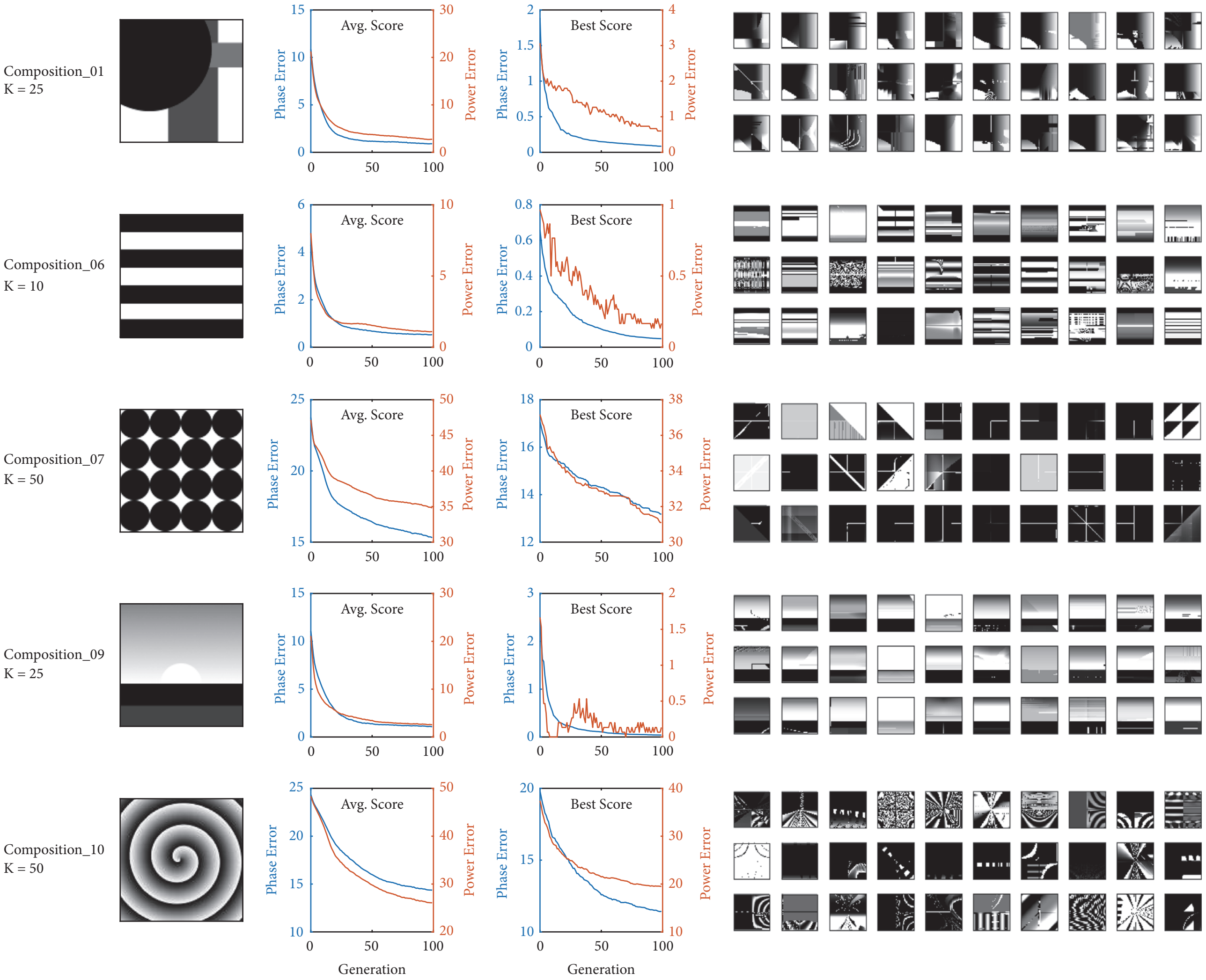


Figure 7: Compositional summary charts and examples. Each row of the figure captures the summary for a target over 30 runs. Leading with the target image, the next two columns show performance plots of the fitness measure (average over 30 runs). The leftmost plot displays the performance through the population average, where the rightmost plot shows the performance of the best individual of the generation. Aside the plots are the best candidate images produced at termination for each run.

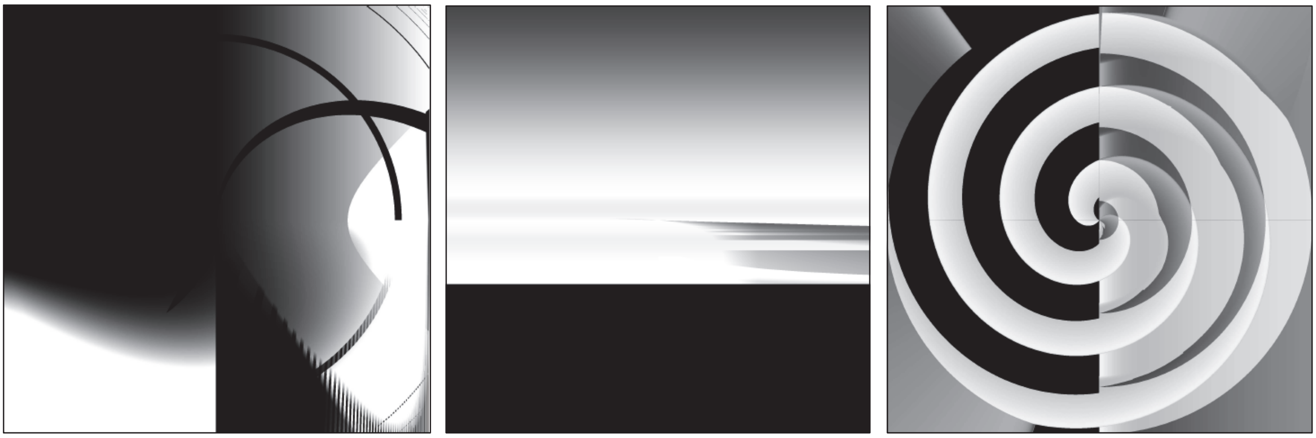
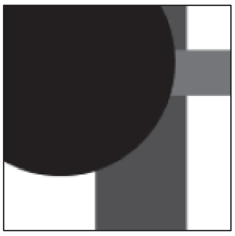
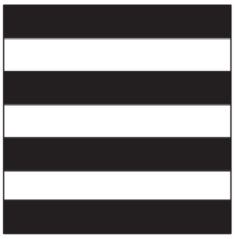
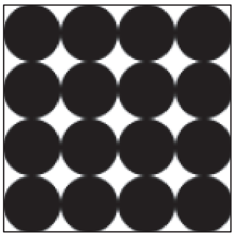
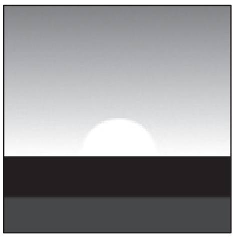
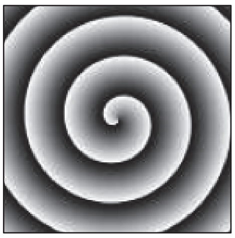


Figure 8: Compositional experiment highlights. Images were produced using targets (from left to right): Composition_01, Composition_09, and Composition_10. Choice of K is outlined in Table 4. These examples show fair replication of compositional aspects, including placement of positions of intensity, contrasts and gradients, and shape characteristics.

Table 4: Compositional fitness summary table. Summaries for the remainder of the section were produced over 30 runs. For each target, a row is included for the mean and standard deviation for each fitness objective aggregate. Our experiment held two objectives, aiming to minimize error in power and phase coefficient matching. The row for "mean" shows the mean of terminal populations' average fitness and the mean of terminal populations' best found candidates across 30 runs. Maximum and expected error values are constrained by choice of K.

	Target	K	Agg.	Power		Phase	
				Mean	Best	Mean	Best
	Composition_01	25	Mean	2.34	0.40	0.79	0.08
			StdDev	0.68	0.81	0.18	0.05
	Composition_06	10	Mean	1.23	0.23	0.53	0.05
			StdDev	0.38	0.43	0.17	0.05
	Composition_07	50	Mean	34.36	30.47	14.31	12.09
			StdDev	3.26	4.44	1.16	1.46
	Composition_09	25	Mean	2.79	0.13	1.17	0.08
			StdDev	0.78	0.51	0.36	0.13
	Composition_10	50	Mean	15.79	9.50	8.51	5.51
			StdDev	4.82	4.66	2.89	3.04

evolutionary art applications. We then evaluate some possible multiobjective adaptations of our measures, and expand our capabilities from grayscale to coloured textures across multiple colour schemes. Finally, we present a brief discussion which corroborates a related measure in previously published research relating to computational aesthetics.

6.1. Language and Representation

6.1.1. Polar Coordinates, Geometric Operators. The first adjustment to our GP language was motivated by the poor performance observed when using targets which displayed strong radial attributes. We found that the inclusion of polar

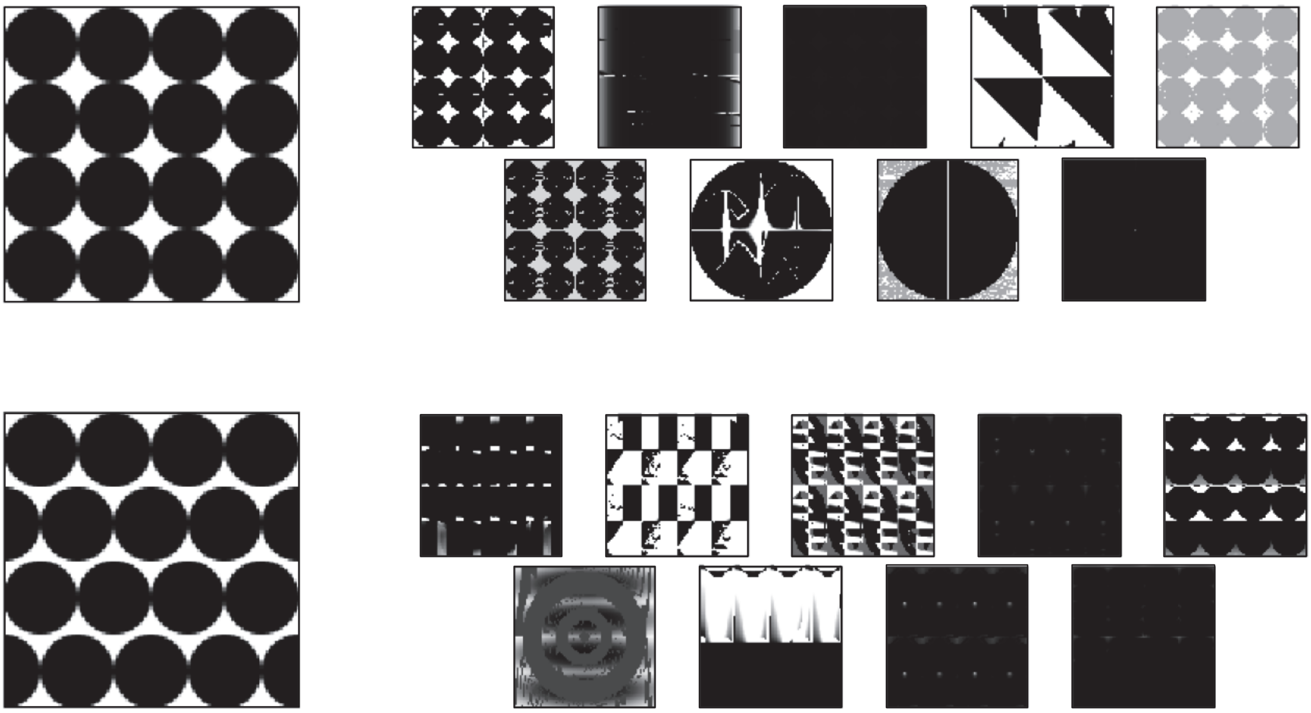


Figure 9: Circle, grid, and offset language summary examples.

coordinate variables improved results for certain compositional targets (e.g., spirals) and some artistic genre targets. Some difficulty was still found with other targets using radial variations and repetitions. We therefore added a set of GP language operators well-suited for these target images.

With inspiration from the Gentropy system by Weins [50], we included the *circle* geometric operator (which returns 1.0 if the current texel is within the provided radius from the origin), along with the coordinate operators of *tile* and *shift*. The *circle* operator provided a simplified way for the candidate programs to show hard transitions about a radius, and the *tile* operator provided an easy way to create arbitrary $n \times m$ tilings.

Figure 9 shows a much-improved set of evolved candidate textures over our previous experiments. We see the error for these two targets decreases by ~40% in both objectives, and a 2-sample t-test provides at most $p < 0.0001$ across objectives and targets, suggesting fair statistical significance when considered with the reduced run count. The performance gains seen with these additional language operators is another promising sign for our fitness measure, and reinforces the importance for GP texture language adequacy.

6.1.2. Noise Generation. To help generate images having more visual complexity and interest, we included numerous noise generation operators (Section 4.2.2). With regard to error values, the introduction of the noise operators appears to be an improvement for most targets. We find minor but consistent reductions in both phase and power errors.

Figure 10 highlights some of the finer details in a pair of larger renderings using a target photograph of a flower.

6.1.3. Coordinate Variable Reduction. One final language experiment was performed by removing the X coordinate variable from the language set. It was expected that removing a fundamental coordinate variable would result in substantial difficulty for our system to produce results, and consequently, high error scores.

It is surprising to see that, despite the previous problems encountered while lacking the polar coordinate variables, there were few noted changes to performance. For the compositional target set, most targets performed only slightly better numerically with the inclusion of the X coordinate, and no statistical significance was found to favour either language set.

When we inspect the evolved textures a little more closely, there appears to be two main ways that our system and its textures adapted to the missing coordinate variable. Some candidates were able to glean sufficient positional information from the remaining coordinate variables: Y, ρ, and ϕ.

An alternative approach appears to largely forgo any direct positional information and instead builds upon layering multiple noise operators. We see this with the highlighted flower images in Figure 11, and a particularly interesting example of the Van Gogh target in Figure 12.

6.2. Colour. Here we considered the approach of evolving colour textures through separate evaluation of each colour channel, along with evaluation across average luminance. Further experimentation with HSL colour models, and other colour analyses can be found in [26].

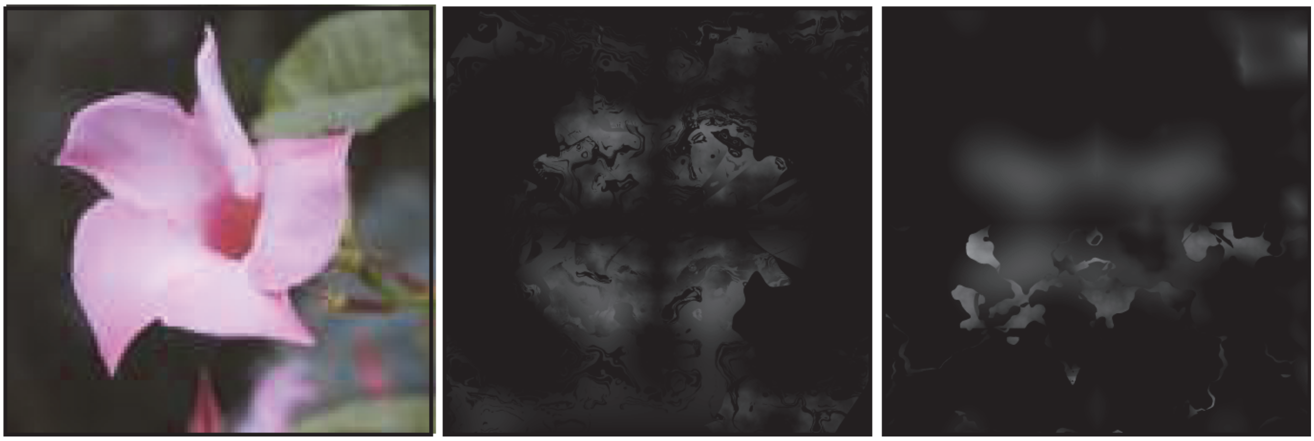


Figure 10: Produced image highlights; noisy language; flower. The leftmost image shows the target, followed by a pair of notable evolved candidates.

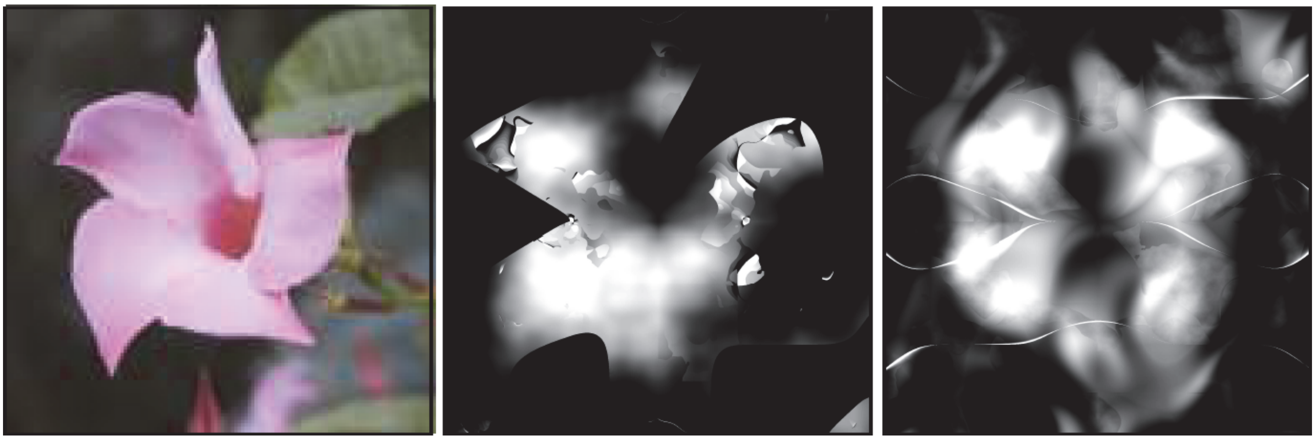


Figure 11: Produced image highlights; noisy language, no X; flower. The leftmost image shows the target, followed by a pair of notable evolved candidates.

We maintained the selection of $K = 50$ as it produced suitable compositional results. To produce colour images, we evolved three GP trees per individual, corresponding to the RGB colour channels. With the increased tree count, and proportional increase in rendering complexity, we performed 9 runs per target. The system was then given 8 fitness objectives to optimize: the original grayscale power (Y) and phase (Y), colour power (R, G, B), colour phase (R, G, B).

6.2.1. Y+RGB Colour Channels. As we found success with our existing measure on grayscale textures, we expanded upon this as a base. The placement and proportion of specific colours is guided using the same measurement technique across each individual RGB colour channel. Where a grayscale texture had two objectives (power and phase), our 4-channel (Y+RGB) colour image used $2 \times 4 = 8$ objectives. Each channel was evaluated similarly to a separate grayscale texture.

We maintained the use of a luminance channel evaluation as it was expected to further constrain the overall composition of the image. It was also hoped that the luminance channel could capture some spatial information lost by assessing colour channels in isolation. We hypothesized that including this combination of luminance and colour channel objectives should reduce attempts to sacrifice any individual colour channel objective by incurring further penalties from mean luminance degradation. The NTSC (CCIR 601) method was used for conversion from colour (RGB) to grayscale:

$$Y = 0.299R + 0.587G + 0.114B \tag{8}$$

This provided a close approximation of colorimetric luminance from the nonlinear, gamma corrected RGB values.

The results in Figure 13 show that the control of colour through relative proportion and overlay of RGB channels, while basic and limited, is successful with certain targets. From the charting, we see similar sacrifices being made to the blue channel power error on target Composition_15. For Composition_14-15, the green channel, while still worse than when evolved in monochrome, sees some slight improvements. Composition_13 sees an overall improvement to shape, where Composition_01 remains consistent.

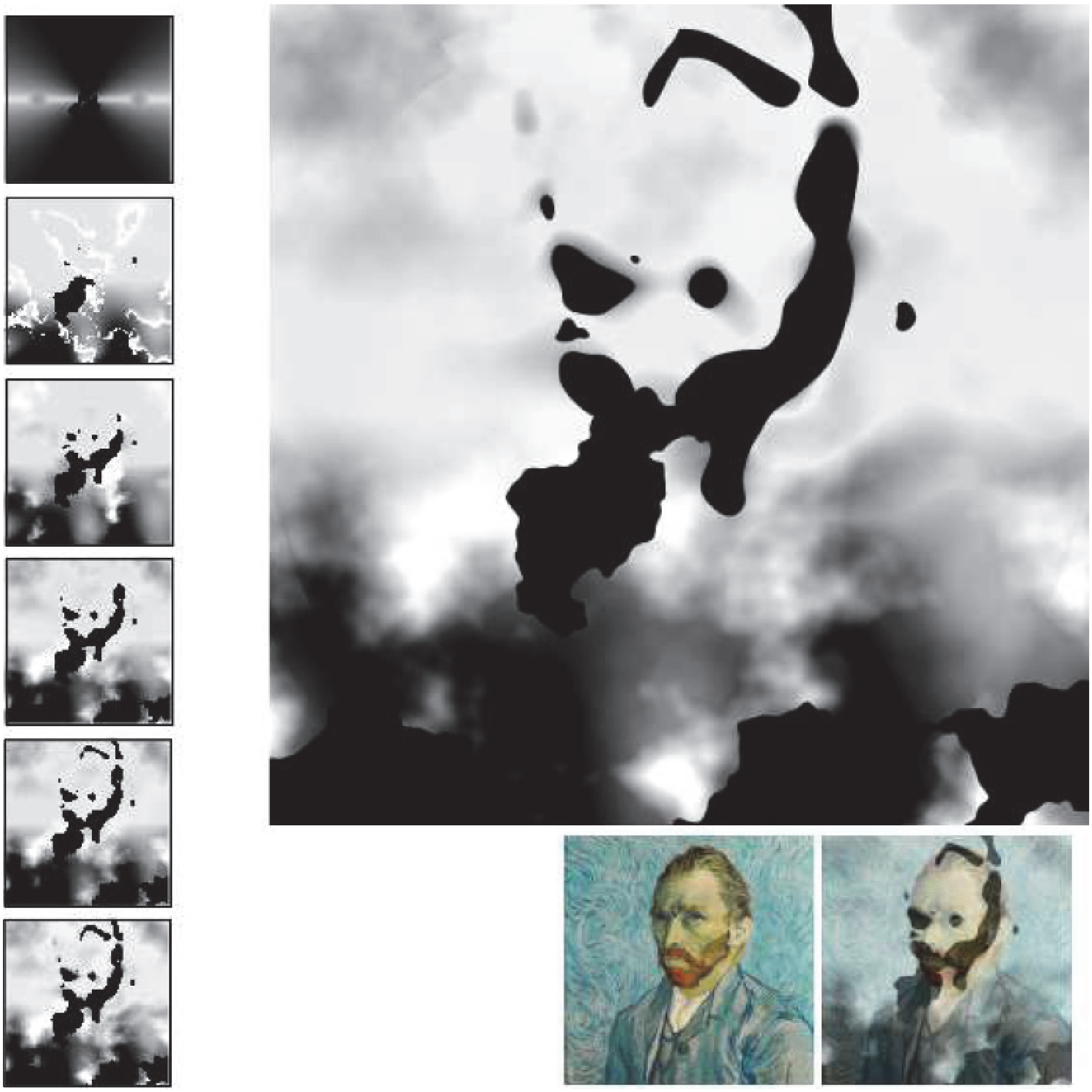


Figure 12: Produced image highlights; noisy language, no X; Van Gogh. On the left, we can see a snapshot of the candidate at every 20 generations. An evolutionary strategy has emerged which gradually applies layers and refines noise operators. The candidate is viewed atop the target image with partial transparency in the bottom left.

While we had hoped that the inclusion of a luminance channel would reduce the occurrence of sacrificing individual colour objectives, we occasionally see the opposite. There is now further pressure to sacrifice an objective if its channel is not contributing positively to the compositional shape as viewed through the lens of averaged luminance.

While overall colour distribution could be improved, we see increased performance when targets hold colour channels which can be replicated as grayscale targets individually. While considering the limitations, we are still able to replicate variations of shape and colour for a number of targets. Some highlights have been shown in Figures 14, 15, and 16.

6.3. Spatial Frequencies and Comfort. In the course of evolving the many candidate images with each target and experiment set, we identified a number of evolved images which we found unpleasant or uncomfortable to view (see Figure 17). Previous research from Fernandez and Wilkins [3] found correlations between intensity level contrasts at certain spatial frequencies with increased levels of discomfort. We direct readers to their paper for an excellent example of the "uncomfortable property".

The concept of spatial frequency denotes a cyclical nature across a measured space, such as the reoccurrence of Gabor and grating peaks along the width of an image. Our study is predicated over power coefficient positions directly relating to these spatial frequencies. While we found great utility in comparing spatial frequencies relative to image width, human perception requires consideration of an observers field of view. To better capture this, we can use calculations

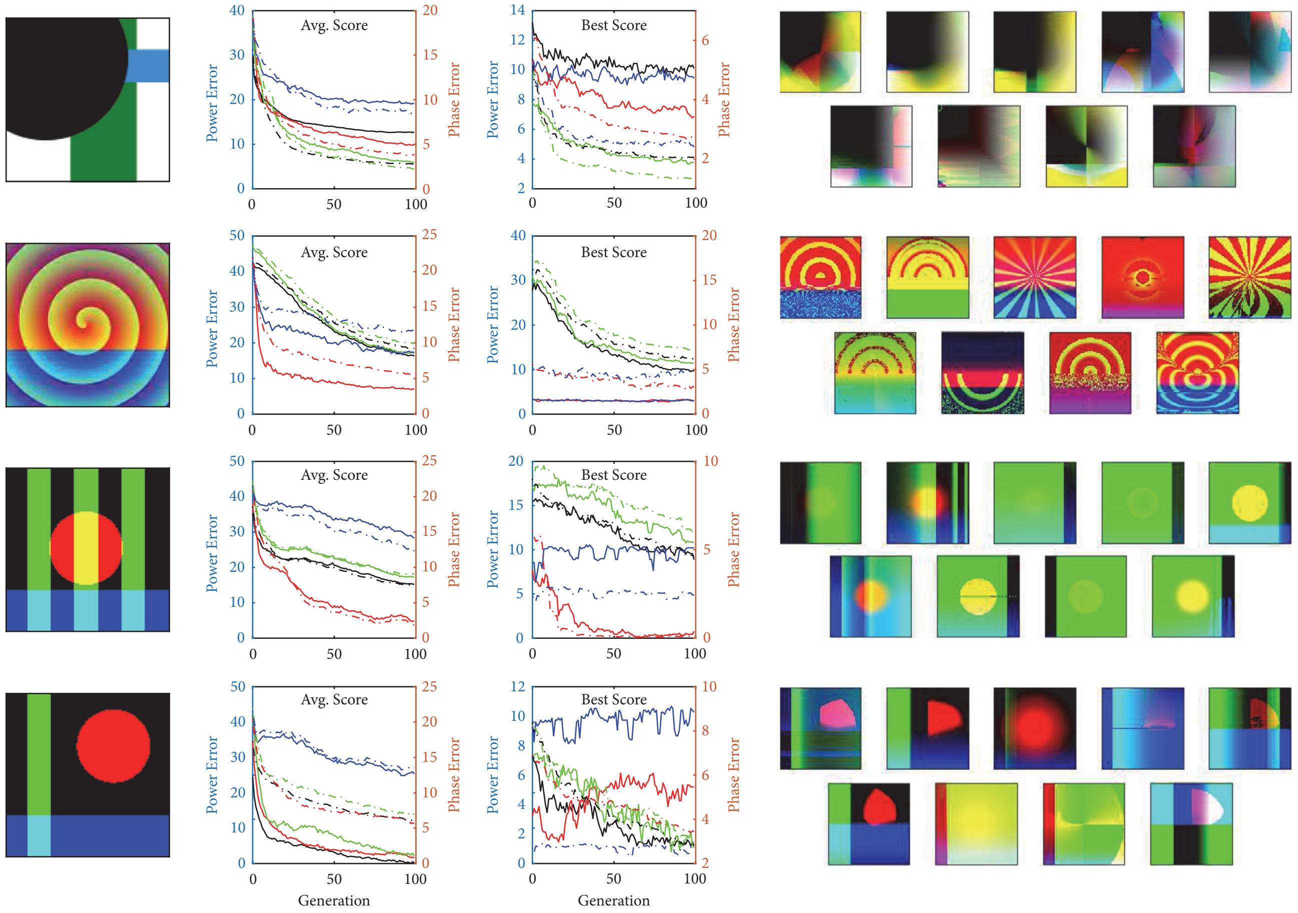


FIGURE 13: Colour experiment summary charts and examples. Power and phase errors for each of the red, green, and blue colour channels are plotted with their respective colour. Errors considered across the average luminance have been plotted in black. Power error is denoted with solid lines, where phase error uses a dashed line.

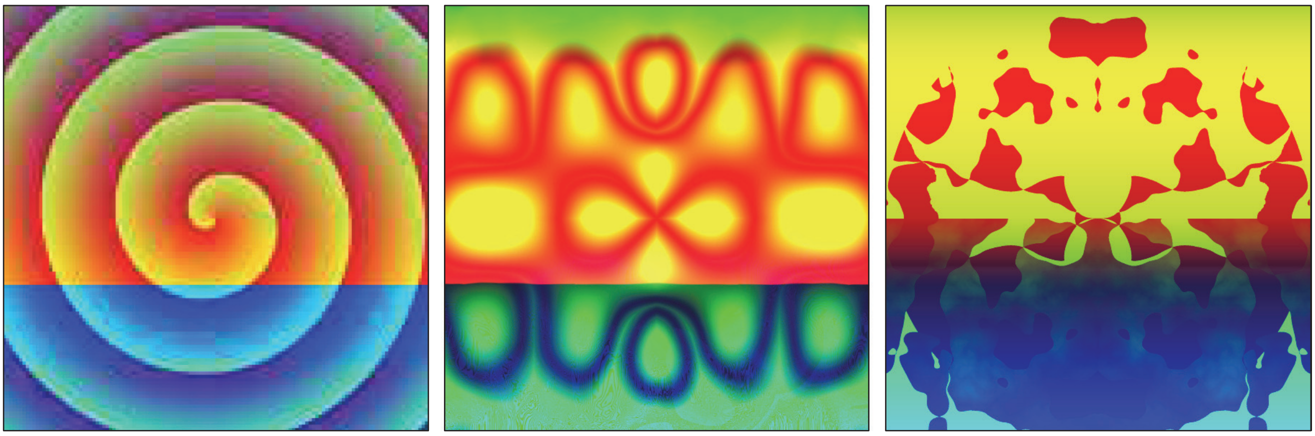


FIGURE 14: Colour experiment highlights; Composition_13; noise language. The leftmost image shows the target, followed by a pair of notable evolved candidates.

of visual angle – when paired with known viewing distance and image size – to compute a relative measure of angular spatial frequency. With spatial frequencies known in relation to image width, we can interpolate their corresponding visual angle when observed with known size and view distance.

Fernandez and Wilkins observed that images with increased amplitudes at a few octaves around 3 cycles per visual degree corresponded with higher reports of image discomfort. We explored numerous schemes in the previous sections to constrain and obtain specific spatial frequencies

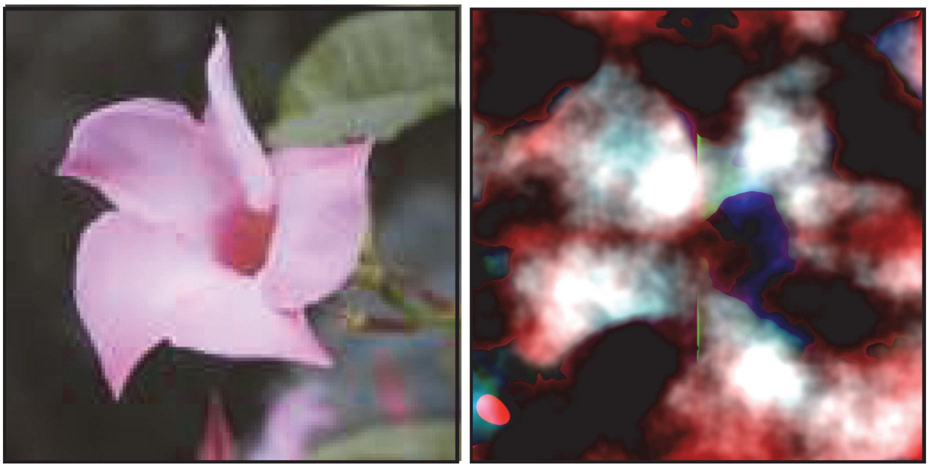


Figure 15: Colour experiment highlights; flower; noise language. The leftmost image shows the target, followed by a notable evolved candidate.

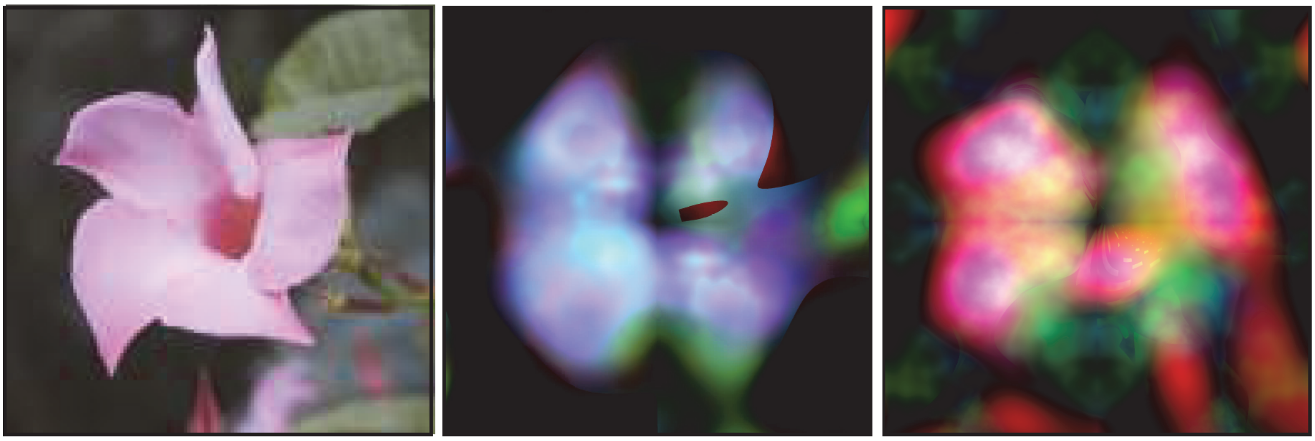


Figure 16: Colour experiment highlights; flower; noise language, no X. The leftmost image shows the target, followed by a pair of notable evolved candidates.

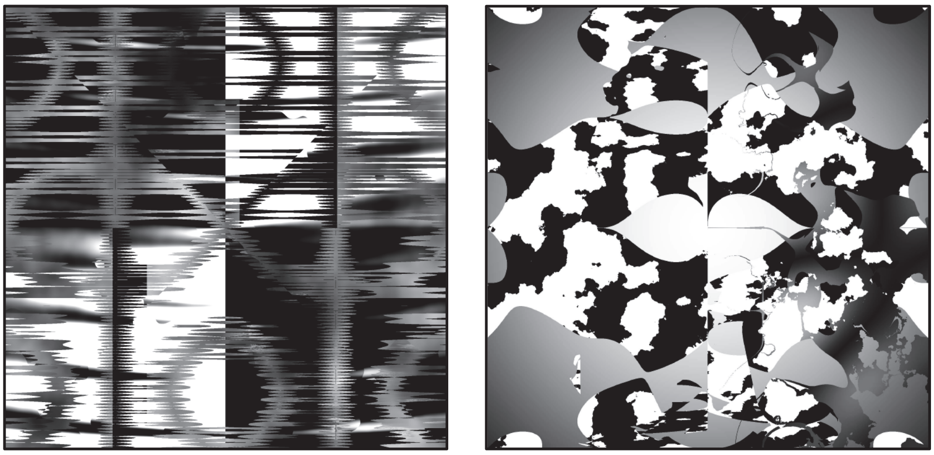


Figure 17: Evolved images with uncomfortable spatial properties. Selection of images with uncomfortable aspects was performed with images sized to 12” side lengths at a viewing distance of 24” (identical angular spatial frequencies can be obtained when this page is viewed at a distance of 4.1”, though we suspect that the eye strain induced from close proximity viewing will cause further undue discomfort).

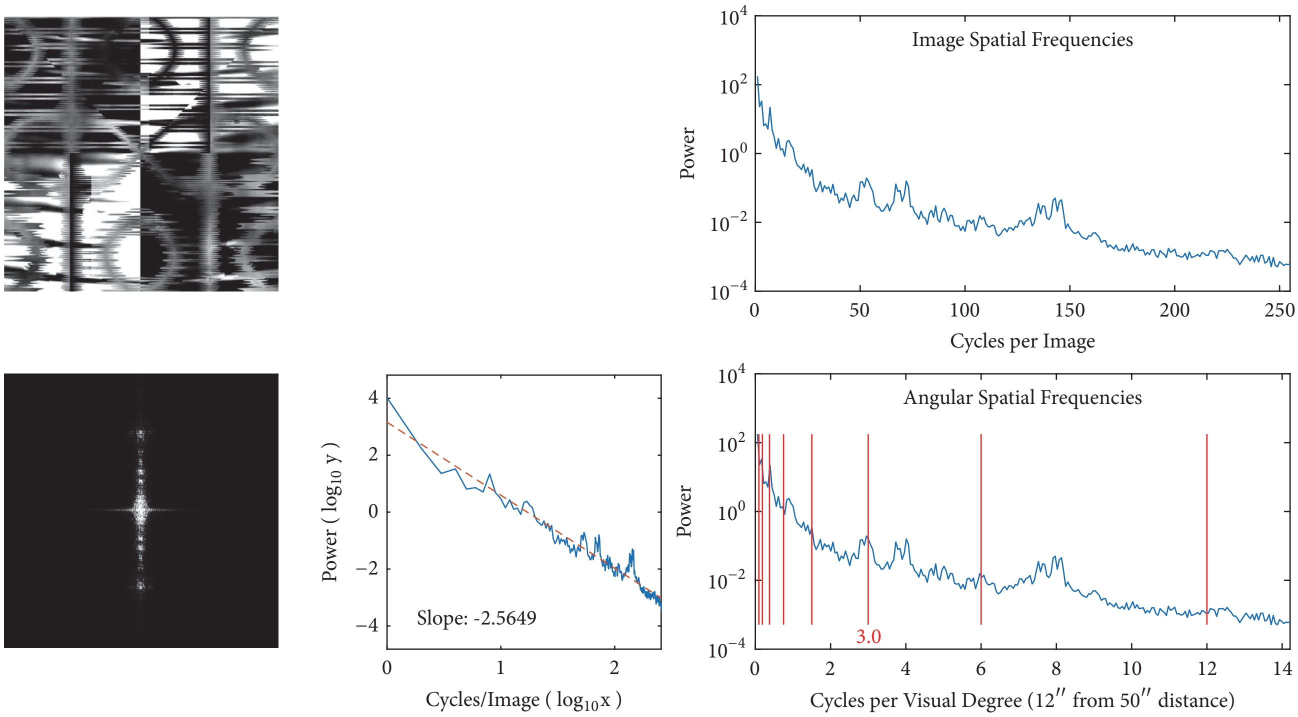


Figure 18: Angular spatial frequency analysis, distance variations. At the top left we have the analysed image, beneath which there are power spectra coefficients display and radially averaged power spectra. The top right graph plots power of the absolute spatial frequencies relative to image width, below which there are the spatial frequencies calculated relative to visual angle at a specific viewing distance. As recommended by Fernandez and Wilkins [3], octaves about 3 cycles/degree have been marked.

of a target image in our newly evolved candidates. With a direct relation between relative visual degree and absolute image spatial frequencies, we posit that these findings may be combined to the effect of a new aesthetic model.

Spatial frequency theory proposes that the human visual cortex operates through analysis of light receptor spatial frequencies [21, 22]. With supporting works finding sensitivity in animals to certain spatial frequency ranges [25, 51], it is not surprising to think that humans may also be more sensitive to contrasts at certain spatial frequencies. As seen in Figure 17, we can corroborate that intensity contrasts at certain visual frequencies are uncomfortable, discordant, and at times even painful. Figure 18 shows a frequency analysis of one of these evolved images. As can be seen, there is a peak in amplitude at the 3 cycles/degree frequency identified by Fernandez and Wilkins [3]. However, this measurement is dependent upon the viewing distance to the image, and this peak at 3 cycles/degree changes with different viewing distances.

With these findings, we identify a couple of limitations in using frequency for the analysis of uncomfortable images. The first, and least negotiable concern, holds that viewing size and distance must be considered before evolution. With interactive or hybrid fitness depending on user-evaluated thumbnails, large incongruities may appear between the rated thumbnails and full-size renderings.

There is another critical concern, though one we are now most capable of identifying and accommodating: naïve reduction to power within a range of frequencies can alter an image to something unrecognisable. We have seen that core compositional information can be stored in 50 or so positions, as witnessed with our experiments in choice for truncation size, K. We can easily expect some of these critical frequencies to lay within the "3.0± two octaves" range identified, and so a blanket frequency reduction should expect poor results with spatial similarity. If no other spatial attributes are sought in the evolved images, this penalty for power in the 3.0 angular spatial frequency range could provide a novel aesthetic measure for exploration. Some refinements will be needed otherwise. If provided target power spectra, we might propose an aesthetic objective which penalizes a surplus of power in these frequency ranges. From our observations above, we might also suggest a distribution of weights to provide harsher penalties when closer to the 3.0 cycles/degree mark.

Despite a number of concerns having been identified, our exploration with power spectra fitness measures has given us a tool to resolve some of them. We also suspect that beyond the correlation with discomfort and the given angular spatial frequency ranges, there may be a need to consider interactions with the phase of these frequencies and their harmonics. With further exploration in the future, novel aesthetic models can be developed from these findings.

7. Conclusion

2D power spectra can be an effective tool for guiding the evolutionary synthesis of images. By applying a 2D Fourier analysis of a target image, key spatial characteristics can be extracted from it and used as a guide for the evolution of images that share these characteristics. Precise duplication of a target image is not desirable. Rather, by focussing on the major frequencies and their spatial orientations, the evolutionary art system is given enough freedom to "fill in the gaps" and generate interesting variations of images that have visual relationships to a target. Thus the approach acknowledges one of the strengths of evolutionary art, and evolution in general: the ability to generate creative and interesting solutions to problems.

Another unexpected result is the possible application of power spectra in identifying evolved images which have uncomfortable properties. A few example images show the spectral properties previously identified by Fernandez and Wilkins [3] in their study of uncomfortable art. Although more research on this topic is needed, there is the possibility of using such analyses within fitness strategies in order to avoid production of images with undesirable visual properties.

The success of the results shown in this paper depends upon two key factors. First, our coefficient reduction scheme proves effective in refining the search by simplifying the computational optimizations required in reproducing Fourier coefficients. Although further improvements and enhancements to this strategy are possible, our approach is generally effective for compositional targets and produced the results shown. Second, it is important that the procedural texture language used in the GP system has adequate power for producing images that conform to characteristics seen in the target image. The property of language adequacy and bias is well known in GP research. With our system, some target images are trivial to reproduce, where others are consistently difficult to handle with the basic procedural texture language. Improvements immediately arise when the language is supplemented with polar coordinates, noise generators, tiling operators, or other language features as needed by the target. On the other hand, some photographs we used as target images rarely yield successful outcomes, even with these additions. We hypothesize that our texture language remains incapable of easily generating images that match these targets. An enhanced texture language and coefficient reduction scheme may be warranted in these more challenging cases.

In summary, computer vision strategies such as spectral analysis continue to show wide success in applications involving image analysis, art classification, image retrieval, and other applications. These techniques should be given serious consideration in evolutionary art as well, in order to improve the quality and sophistication of machine-synthesized art.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by NSERC Discovery Grant [RGPIN-2016-03653].

References

[1] P. Bentley and D. W. Corne, *Creative Evolutionary Systems*, Morgan Kaufmann, 2002.

[2] J. Romero and P. Machado, *The Art of Artificial Evolution*, Springer, Berlin, Heidelberg, 2008.

[3] D. Fernandez and A. J. Wilkins, "Uncomfortable images in art and nature," *Perception*, vol. 37, no. 7, pp. 1098–1113, 2008.

[4] D. J. Graham, "Art statistics and visual processing: insights for picture coding," in *Proceedings of the Picture Coding Symposium, PCS '09*, pp. 1–4, IEEE, May 2009.

[5] D. Neumann and K. Gegenfurtner, "Image Retrieval and Perceptual Similarity," *ACM Transactions on Applied Perception*, vol. 3, no. 1, pp. 31–47, 2006.

[6] W. Niblack, R. Barber, W. Equitz et al., "Qbic project: querying images by content, using color, texture, and shape," in *Proceedings of the IS and T/SPIE's Symposium on Electronic Imaging: Science and Technology*, pp. 173–187, International Society for Optics and Photonics, February 1993.

[7] D. S. Ebert, *Texturing and Modeling: A Procedural Approach*, Morgan Kaufmann, 2003.

[8] M. Hull and S. Colton, "Towards a general framework for program generation in creative domains," in *Proceedings of the 4th International Joint Workshop on Computational Creativity, IJWCC '07*, pp. 137–144, June 2007.

[9] M. Baniasadi and B. J. Ross, "Exploring non-photorealistic rendering with genetic programming," *Genetic Programming and Evolvable Machines*, vol. 16, no. 2, pp. 211–239, 2015.

[10] C. Reynolds, "Interactive evolution of camouflage," *Artificial Life*, vol. 17, no. 2, pp. 123–136, 2011.

[11] M. Hendrikx, S. Meijer, J. Van Der Velden, and A. Iosup, "Procedural content generation for games: a survey," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 9, no. 1, pp. 1–22, 2013.

[12] J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, University of Michigan Press, 1975.

[13] K. Sims, "Artificial evolution for computer graphics," *ACM SIGGRAPH Computer Graphics*, vol. 25, no. 4, pp. 319–328, 1991.

[14] J. R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, MIT Press, 1992.

[15] S. Baluja, D. Pomerleau, and T. Jochem, "Towards automated artificial evolution for computer-generated images," *Connection Science*, vol. 6, no. 2-3, pp. 325–354, 1994.

[16] A. E. M. Ibrahim, *Genshade: an evolutionary approach to automatic and interactive procedural texture generation [Ph.D. thesis]*, Texas A and M University, 1998.

[17] A. L. Wiens and B. J. Ross, "Gentropy: evolving 2D textures," *Computers & Graphics*, vol. 26, no. 1, pp. 75–88, 2002.

[18] P. Dahlstedt, "Turn-based evolution as a proposed implementation of artistic creative process," in *Proceedings of the 2012 IEEE Congress on Evolutionary Computation, CEC '12*, pp. 1–7, IEEE, June 2012.

[19] D. J. Graham and D. J. Field, "Variations in intensity statistics for representational and abstract art, and for art from the Eastern

and Western hemispheres," *Perception*, vol. 37, no. 9, pp. 1341–1352, 2008.

[20] D. J. Graham and C. Redies, "Statistical regularities in art: Relations with visual coding and perception," *Vision Research*, vol. 50, no. 16, pp. 1503–1509, 2010.

[21] M. B. Sachs, J. Nachmias, and J. G. Robson, "Spatial-frequency channels in human vision," *Journal of the Optical Society of America*, vol. 61, no. 9, pp. 1176–1186, 1971.

[22] L. Maffei and A. Fiorentini, "The visual cortex as a spatial frequency analyser," *Vision Research*, vol. 13, no. 7, pp. 1255–1267, 1973.

[23] P. Vuilleumier, J. L. Armony, J. Driver, and R. J. Dolan, "Distinct spatial frequency sensitivities for processing faces and emotional expressions," *Nature Neuroscience*, vol. 6, no. 6, pp. 624–631, 2003.

[24] A. Fiorentini and N. Berardi, "Perceptual learning specific for orientation and spatial frequency," *Nature*, vol. 287, no. 5777, pp. 43-44, 1980.

[25] R. L. De Valois, D. G. Albrecht, and L. G. Thorell, "Spatial frequency selectivity of cells in macaque visual cortex," *Vision Research*, vol. 22, no. 5, pp. 545–559, 1982.

[26] M. Gircys, *Image evolution using 2d power spectra [M.S. thesis]*, Brock University, 2018.

[27] J. M. Brayer, "Introduction to fourier transforms for image processing," https://www.cs.unm.edu/brayer/vision/fourier.html.

[28] P. Brémaud, *Mathematical Principles of Signal Processing: Fourier and Wavelet Analysis*, Springer Science and Business Media, 2013.

[29] A. Rauh and G. R. Arce, "Sparse 2d fast fourier transform," in *Proceedings of the 10th International Conference on Sampling Theory and Applications*, 2012.

[30] S. J. Sangwine, "The problem of defining the fourier transform of a colour image," in *Proceedings of the International Conference on Image Processing, ICIP '98*, vol. 1, pp. 171–175, IEEE, October 1998.

[31] V. R. Dubey, "Quaternion fourier transform for colour images," *International Journal Computer Science and Information Technology*, vol. 5, no. 3, 2014.

[32] National Semiconductor, "Power spectra estimation," http://www.dcs.warwick.ac.uk/feng/teaching/PowerSpectrum.pdf.

[33] C. E. Jacobs, A. Finkelstein, and D. H. Salesin, "Fast multiresolution image querying," in *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, pp. 277–286, ACM, 1995.

[34] Y. Li, C. Hu, L. L. Minku, and H. Zuo, "Learning aesthetic judgements in evolutionary art systems," *Genetic Programming and Evolvable Machines*, vol. 14, no. 3, pp. 315–337, 2013.

[35] B. J. Ross and H. Zhu, "Procedural texture evolution using multi-objective optimization," *New Generation Computing*, vol. 22, no. 3, pp. 271–293, 2004.

[36] E. Den Heijer and A. E. Eiben, "Investigating aesthetic measures for unsupervised evolutionary art," *Swarm and Evolutionary Computation*, vol. 16, pp. 52–68, 2014.

[37] P. Machado and A. Cardoso, "Computing aesthetics," in *Proceedings of the XIVth Brazilian Symposium on AI*, pp. 239–249, Springer, Berlin, Germany, 1998.

[38] P. Machado and A. Cardoso, "All the truth about NEvAr," *Applied Intelligence*, vol. 16, no. 2, pp. 101–118, 2002.

[39] B. J. Ross, W. Ralph, and H. Zong, "Evolutionary image synthesis using a model of aesthetics," in *Proceedings of the 2006 IEEE Congress on Evolutionary Computation*, G. G. Yen, L. Wang, P. Bonissone, and S. M. Lucas, Eds., pp. 3832–3839, IEEE Press, Vancouver, Canada, July 2006.

[40] F. R. Tanjil, *Deep learning concepts for evolutionary art [M.S. thesis]*, Brock University, 2018.

[41] B. Julesz and T. Caelli, "On the limits of Fourier decompositions in visual texture perception," *Perception*, vol. 8, no. 1, pp. 69–73, 1979.

[42] The MathWorks, Inc., Mathworks: Matlab, https://www.mathworks.com/.

[43] S. Luke, L. Panait, G. Balan et al., "Ecj: A java-based evolutionary computation research system," 2006, http://cs.gmu.edu/eclab/projects/ecj.

[44] Unported http://creativecommons.org/licenses/by/3.0/]., "Riven, Perlinnoise: smooth/turbulent," Creative Commons Attribution 3.0, http://riven8192.blogspot.ca/2009/08/perlinnoise.html.

[45] S. Gustavson, "Simplex noise demystified," http://staffwww.itn.liu.se/stegu/simplexnoise/simplexnoise.pdf.

[46] C. A. C. Coello, G. B. Lamont, and D. A. Van Veldhuizen, *Evolutionary Algorithms for Solving Multi-Objective Problems*, Kluwer Academic, New York, NY, USA, 2nd edition, 2007.

[47] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison Wesley, 1989.

[48] P. J. Bentley and J. P. Wakefield, "Finding acceptable solutions in the pareto-optimal range using multiobjective genetic algorithms," in *Soft Computing in Engineering Design and Manufacturing*, Springer, Berlin, Germany, 1997.

[49] D. W. Corne and J. D. Knowles, "Techniques for highly multiobjective optimisation: Some nondominated points are better than others," in *Proceedings of the 9th Annual Genetic and Evolutionary Computation Conference, GECCO '07*, pp. 773–780, ACM Press, July 2007.

[50] A. L. Wiens and B. J. Ross, "Gentropy: Evolutionary 2D texture generation," in *Proceedings of the Late Breaking Papers at the 2000 Genetic and Evolutionary Computation Conference*, D. Whitley, Ed., pp. 418–424, July 2000.

[51] N. P. Issa, C. Trepel, and M. P. Stryker, "Spatial frequency maps in cat visual cortex," *The Journal of Neuroscience*, vol. 20, no. 22, pp. 8504–8514, 2000.

Hindawi
Complexity
Volume 2018, Article ID 9270152, 16 pages
https://doi.org/10.1155/2018/9270152

Research Article

Evolutionary Computation for Modelling Social Traits in Realistic Looking Synthetic Faces

Felix Fuentes-Hurtado, Jose A. Diego-Mas, Valery Naranjo, and Mariano Alcañiz

i3B, Institute for Research and Innovation in Bioengineering, Universitat Politecnica de Valencia, 46022 Valencia, Spain

Correspondence should be addressed to Jose A. Diego-Mas; jodiemas@dpi.upv.es

Received 8 June 2018; Revised 7 September 2018; Accepted 23 September 2018; Published 23 October 2018

Guest Editor: Juan Romero

Human faces play a central role in our lives. Thanks to our behavioural capacity to perceive faces, how a face looks in a painting, a movie, or an advertisement can dramatically influence what we feel about them and what emotions are elicited. Facial information is processed by our brain in such a way that we immediately make judgements like attractiveness or masculinity or interpret personality traits or moods of other people. Due to the importance of appearance-driven judgements of faces, this has become a major focus not only for psychological research, but for neuroscientists, artists, engineers, and software developers. New technologies are now able to create realistic looking synthetic faces that are used in arts, online activities, advertisement, or movies. However, there is not a method to generate virtual faces that convey the desired sensations to the observers. In this work, we present a genetic algorithm based procedure to create realistic faces combining facial features in the adequate relative positions. A model of how observers will perceive a face based on its features' appearances and relative positions was developed and used as the fitness function of the algorithm. The model is able to predict 15 facial social traits related to aesthetic, moods, and personality. The proposed procedure was validated comparing its results with the opinion of human observers. This procedure is useful not only for creating characters with artistic purposes, but also for online activities, advertising, surgery, or criminology.

1. Introduction

Since ancient times, people believe that the face is a window to the true nature of a person, the most direct way to their emotions and feelings [1]. People use information from faces to identify others, to guess their gender, age, or race, to make attributions such as personality, intelligence, or trustworthiness [2], or even to judge the emotions and intentions of the owners of the faces [3]. Our brain is specially efficient perceiving faces [4, 5] and processing the information extracted from them. These attributions are formed very fast; 34 milliseconds of exposition is enough for human brain to create a first impression of a face. So, the appearance of faces plays a central role in our everyday decisions [6–8] and in our relationships with other people [9]. For example, voting decisions [6, 10], criminal justice decisions [11, 12], mate selection [13–15], or how we choose social partners [16] is influenced by what we perceive in the face of others.

Faces play a central role in art, design, or advertising to convey and elicit emotions. How a face looks in a painting or an advertisement can dramatically influence what we feel about them and what emotions are elicited. Studies are still being made on the face of the Mona Lisa and the emotions that her face conveys [17]. Previous works have proved that when looking at scenes containing human faces, observers tend to rapidly focus on the faces [18], even if faces do not occupy the most part of the scene. But faces are not important only for arts. Due to the importance of appearance-driven judgements of faces, face perception has become a major focus not only for psychological research, but for neuroscientists, engineers, and software developers [19]. New human-machine interaction systems and online activities like e-commerce, e-learning, games, dating, or social networks are fields in which it is common to use human digital representations that symbolize the user's presence or that act as virtual interlocutor [20]. The importance of communicative behaviours of these avatars in new interaction systems [21–25] has led to an increasing interest in creating realistic virtual faces able to convey appropriated sensations to users [26–29].

The objective of this work is to develop a system to generate realistic looking synthetic faces that transmit to human observers the sensation of having a set of social traits each of them in a preestablished amount. The developed system must create faces with appropriate facial features to achieve this objective. Hereinafter, *social traits* will be used as any judgement that a human observer can make about the aesthetic characteristics of a face (e.g., attractiveness) or about the emotional state (e.g., sadness) or personality (e.g., dominance) of the owner of the face. In the same way, *facial features* will refer to the morphological characteristics of the faces.

Developing such a system must overcome two great difficulties. The first one is to establish the relationships between the facial features of a face and its social traits. Visual perception research has shown that human brain processes faces in different way to other kinds of objects [30]. Part-based perceptual models suppose that objects are processed on the basis of their components or parts [31]; although it is commonly agreed that this is the way in which we process most objects, faces are thought to be processed in a different way. In *relational* [32] or *configural* [33] models of perception, first-order features (like isolated face features) are processed in a part-based way, but second- and higher-order features emerge from the combination of several lower-order features, and these are used to make judgments from faces. The amount of information derived from second- and higher-order features used depends on the kind of judgment that is made from faces [32]. For example, it is suggested that face recognition depends mainly on first-order features and part-based information processing [34, 35], while more complex judgments require information from second- and higher-order features. Holistic perceptual models integrate facial features into a gestalt whole when the human brain processes a face's information (holistic face processing) [36]. The pure holistic processing of faces, with no decomposition into parts, is not supported by the evidences that suggest that some judgements rely mainly on part-based processing of faces [30]. This leads to the mixed holistic/part-based models. These models do not exclude part-based processing from the global holistic processing during face perception [37, 38].

Therefore, to establish the relationships between facial characteristics and social traits elicited in the observers is challenging due to the complexity of the face perception process itself. But, if such a model that relates facial features and social traits is developed, another difficulty remains to create faces that convey a predefined set of social traits. It is possible to consider a face like a set of facial features. This way, the problem is to find the optimal combination of facial features that elicits, simultaneously, a preestablished quantity of each social trait. Therefore, the problem becomes a multiobjective combinatorial optimization problem. Moreover, the number of facial features to be considered can be high (nose, mouth, eyes, eyebrows, relative distances, etc.), as well as the number of possible types of each facial feature (how many types of noses, eyes, jaws, etc.). Therefore, the space of solutions of the problem can be huge.

There are systems to generate realistic synthetic faces and to synthesize emotional facial expressions since the last century [39–42]. A common approach for modelling social traits in artificially generated faces is to systematically modify one facial feature over an existing face, asking people to assess the modified face in the range of the social traits of interest. The modified feature that obtains the best score is fixed and the process is repeated over another facial feature. Considering the holistic face perception model, this approach is far from being optimal. Some other techniques bear in mind that faces are perceived in a gestalt whole rather than as a collection of features independently considered. Among them, two sets of methods can be differentiated: psychological reverse correlation methods (PRCM) and reverse correlation methods in the context of face space models (FSRCM) [3]. PRCM alter faces using randomly generated noise. There are two popular PRCM techniques, both of them consisting in superimposing noise on images. In the first approach, the base face is unambiguous (e.g., a prototypical sad face), while in the second approach, the face is ambiguous (e.g., two facial expressions morphed in one face) [43–45].

While the previous approach made use of noise to achieve its objective, FSRCM approach is focused on changing some characteristics of the faces directly. The procedure can be divided into two tasks: the first one is to develop a model of a face representation, and the second one is to establish the changes in the facial features of the face that lead to the desired changes in social judgments. Similarly to PRCM, FSRCM does not explicitly manipulate facial features. This approach makes use of a faces space, where faces are represented as points in a multidimensional space and each dimension is a property of the face [46]. Oosterhof and Todorov [47] followed this approach to generate models of perceived face trustworthiness, threat, and dominance. In a posterior work, they also built models of several other social traits, such as attractiveness [3, 48]. Walker and Vetter [49] used this procedure for aggressiveness, extroversion, likeability, risk-seeking, social skills, and trustworthiness and used the obtained models to manipulate real faces leading to the expected social attributions.

However, these previous methods have some important limitations. The results of PRCM procedures are models of the strategy used by observers when they assess faces. These models are obtained from a survey in which each participant assesses a big set of artificially degraded faces. The enthusiasm of the participant to perform the task will most likely decay with time, affecting the obtained models [43]. Moreover, both mentioned approaches need a large number of trials to model the expected social attributions in faces, which can lead to lose the participant's motivation and to worsen the quality of the results. Another limitation of reverse correlation methods is that they are limited to create models of one category (e.g., trustworthy, dominant, etc.) per task. Outcomes may change considerably when the objective is to create faces that convey several social traits to some extent, considering simultaneously multiple traits.

In this work, we propose a very different approach to automatically create virtual realistic faces that convey several social traits simultaneously, each of them in a predefined

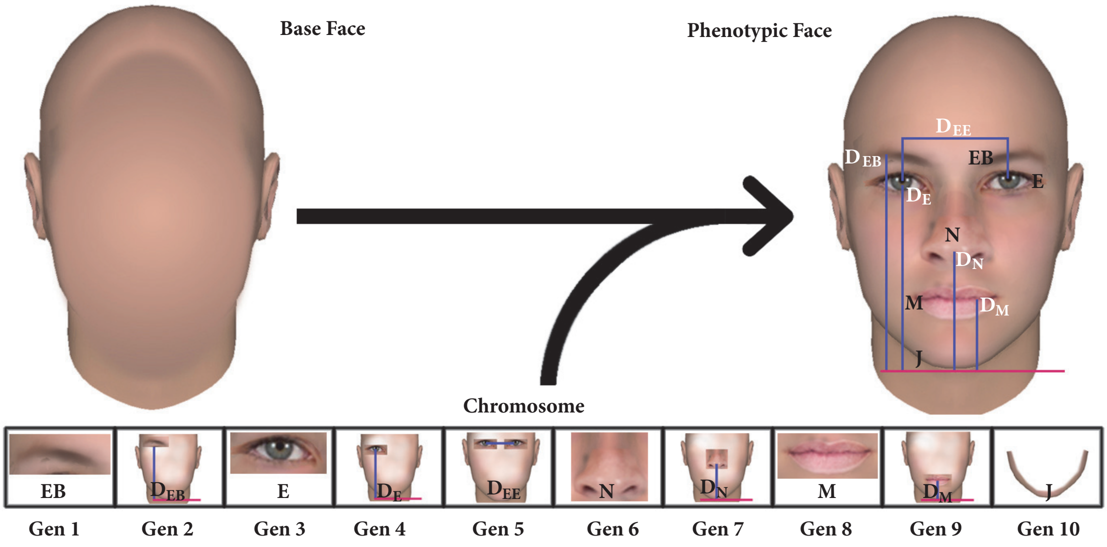


Figure 1: Structure of the chromosomes. A face is constructed by placing on the base face the features indicated by genes 1, 3, 6, 8, and 10, at the positions indicated by genes 2, 4, 5, 7, and 9.

quantity. This approach is, basically, to combine the appropriate set of facial features to form the faces. The facial features and their relative positions must be selected in such a way that impressions elicited in observers were as similar as possible to those established by the designer. In the first step of this approach, an evolutionary algorithm that looks for the adequate set of facial features to elicit the desired social traits is proposed. This kind of algorithms has been used before in evolutionary systems to generate faces of specific identity like EFIT-V [50] or EvoFIT [51].

Secondly, a model that relates the facial features of the faces to the social traits perceived by human observers is developed. This model is used as the fitness function of the evolutionary algorithm. Finally, the optimal set of facial features is combined to shape a realistic looking face. Using this new approach, the designer of the virtual face establishes the amount of each social trait that must be elicited (profile of social traits), and the system automatically generates the proper face.

2. A Genetic Algorithm to Generate Faces

Faces are characterized by their features (two specific eyes, a particular nose, a mouth, etc.) and by the spatial relation between them (relational information). The facial features considered in this work were selected considering previous studies. Internal features (i.e., eyes, nose, and mouth) seem to have significant importance in face recognition [52, 53]. Among the internal features, eyes play a key role in face information processing [54]. Some authors include the eyebrows in the eye area [55, 56] or consider the eyebrows as a major factor in the perception of a face [57]. Blais et al. [58] found that the mouth area is an important cue for both static and dynamic facial expressions, which was consistent with previous researches [59]. However, external facial features such as hair or the shapes of the cheek, the chin, or the jaw also play an important role in the way in which the brain processes the face information. According to Axelrod and Yovel [60], the fusiform face area of the brain is not only sensitive to external features but is also sensitive to their influence on the representation of internal facial features. Some works found that the face shape contributes significantly to faces discrimination [61, 62]. Considering these previous works, we decided to consider the internal facial features (eyebrows, eyes, nose, and mouth) and the jaw contour in this study. Although other features have effect on faces perception, e.g., hair and facial hair, skin tone, and facial proportions [14, 63–67], we limited our study to those features that have a main effect on face perception, rather than considering features that may vary from time to time like hair (people can get a haircut). In addition to these five facial features, the relative positions between them will be considered. D_{EB}, D_E, D_N, and D_M are the vertical positions of the eyebrows, the eyes, the nose, and the mouth, respectively, measured from a horizontal line that passes through the base of the jaw line (Figure 1). D_{EE} is the distance between the centres of the eyes. Therefore, one face can be defined by 10 parameters (EB, E, N, M, J, D_{EB}, D_E, D_N, D_M, and D_{EE}).

The number of faces that can be generated as a combination of these parameters depends on the number of different values that each parameter can take (the number of different eyebrows, noses, mouths, etc.). The number of features of each class included in this study will be discussed later. Considering a minimum of 10 features of each class, the size of the solution space is, at least, 1e10. Due to its complexity, the problem cannot be solved using enumerative or analytic procedures. Therefore, a genetic algorithm (GA) [68, 69] is used to look for the optimal combination of parameters. GAs explore the faces space performing a stochastic guided search based on the evolution of a set (population) of structures

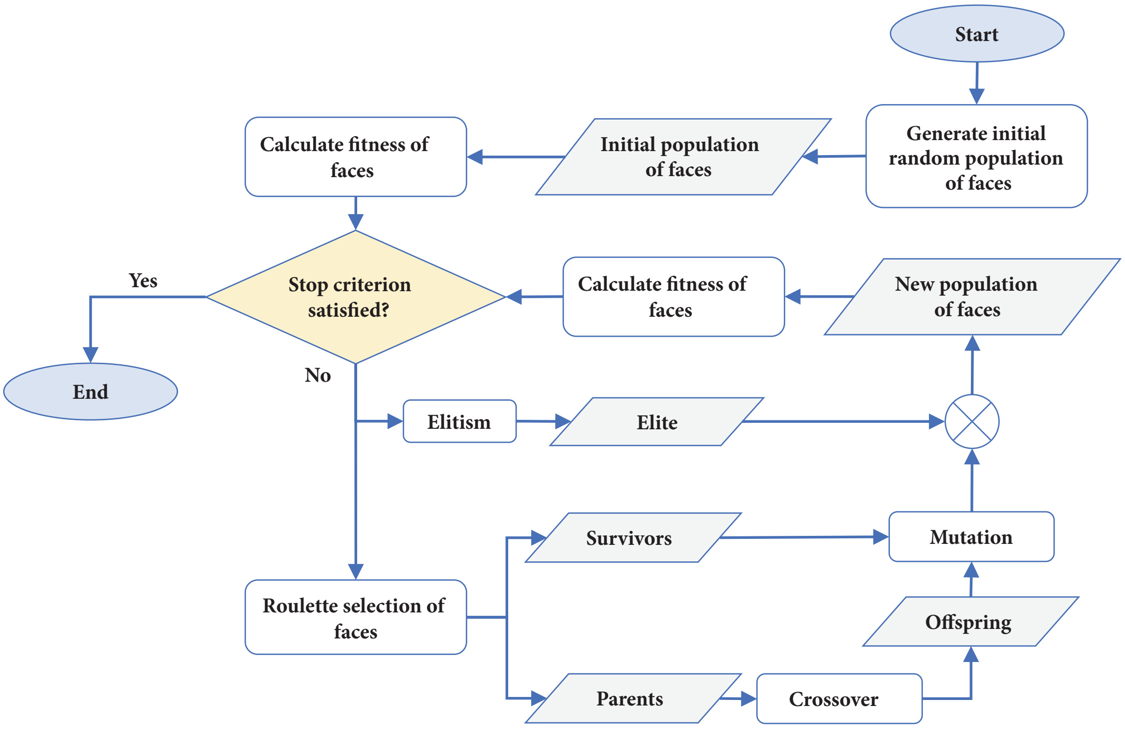

Start
Generate initial random population of faces
Initial population of faces
Calculate fitness of faces
Stop criterion satisfied?
Yes
End
No
Calculate fitness of faces
New population of faces
Elitism
Elite
Survivors
Mutation
Offspring
Roulette selection of faces
Parents
Crossover

Figure 2: Flow chart of the genetic algorithm.

(chromosomes). Each chromosome represents a solution to the problem (a face). The population of faces is evaluated using a fitness function to measure its suitability for the requirements of the problem. Based on the fitness of each chromosome, a new population of faces, which inherit the best characteristics of their predecessors, is obtained. The new population of faces is the result of several transformations guided by genetic operators (selection, crossover, and mutation), which combine or alter the chromosomes obtaining new faces. This iterative procedure is repeated with a predefined number of iterations or until another stop criterion is reached.

Each chromosome is composed of 10 genes (Figure 1). Genes 1, 3, 6, 8, and 10 codify one facial feature of each class. The remaining genes codify the positions in which the features will be located in the face. According to the fundamental theorem of genetic algorithms [69], codifications that favour short and low-order schemata are preferable. Therefore, genes that codify the position of one specific feature have been placed close to the gene that codifies that feature.

The flow chart of the algorithm employed in this work is shown in Figure 2. An initial population of **n** (population size) chromosomes of faces is randomly generated. Roulette wheel selection [68] is used to choose the survivor and reproducer chromosomes in each generation. The ratio between survivors and reproducer is controlled by the $\mathbf{P_c}$ (crossover probability) parameter. The number of survivors is $\mathbf{n}\bullet(1-\mathbf{P_c})-1$, while the number of reproducers is $\mathbf{n}\bullet\mathbf{P_c}$. A single-point crossover operator is used to obtain the offspring from the parents. Mutation operator acts over survivors and the offspring to form a new generation. To complete the **n** chromosomes of the new generation, the best face of the previous generation is always selected to go on to the next (elitism).

The single-point crossover process is shown in Figure 3. After selecting two parents, a crossover point is randomly chosen. Two descendants are produced by merging the genes that remain on each side of the crossover point in each of the parents. The crossover is a closed operator since it always produces chromosomes that represent feasible solutions to the problem. The mutation operator is applied changing the allele that occupies a gene if a random number between 0 and 1 is less than $\mathbf{P_m}$ (mutation probability). The new allele is selected randomly. A typical value for $\mathbf{P_m}$ ranges between 0 and 0.1 [70].

3. A Model to Predict Social Traits Elicited from Facial Features

Two questions remain unsolved in the previously defined evolutionary algorithm. The first one is to establish the alleles of each gene that represent a facial feature in the chromosomes, i.e., the different eyebrows, eyes, noses, mouths, and jaws that will be considered as alleles. The second one is to create a model that relates the facial features that form a face and the social traits perceived by the observers, i.e., the fitness function of the algorithm.

3.1. Alleles of the Facial Features' Genes. The sensations that a face elicits in human observers arise from the visual characteristics of the face. It is not possible to establish the number of different shapes that a human facial feature

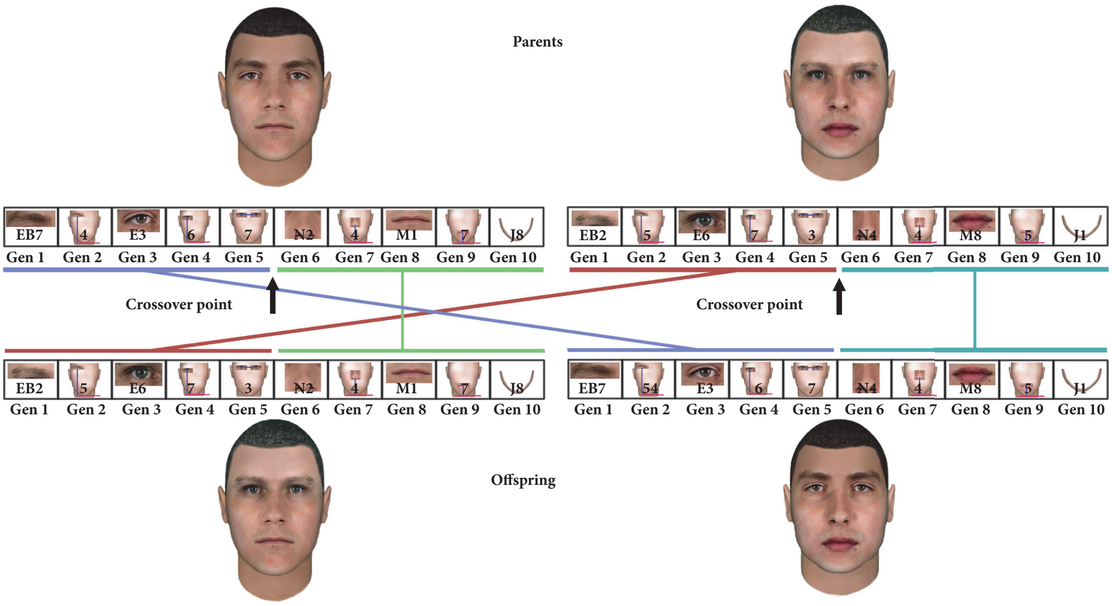


Figure 3: Single-point crossover process. Offspring is obtained by merging the genes on each side of the crossover point in each of the parents.

can take, but it can be supposed that features with similar appearance have the same effect on the perceived social traits. Considering this, we propose to create groups or clusters of features with the same appearance. All the features included in one cluster will elicit very similar sensations in observers. Therefore, all of them can be properly represented by one of the features of this cluster (representative feature). In this way, the number of possible alleles of a gene can be reduced to the number of representative features, i.e., the number of clusters of the feature.

To obtain the features clusters, a set of 93 images of faces (Figure 4(a)) was analysed. After reviewing several well-known databases [71], we selected the Chicago Face Database (CFD) [72]. This database contains high-resolution standardized images of real faces of Asian, Black, Latino, and White males and females with several expressions (including neutral). All the images in the database have the same size and resolution; faces have the same position, pose, and orientation, and the background and illumination are uniform. The homogeneity of the conditions in which the images were obtained was an important factor to select this face database because, for example, differences in the illumination can affect the way in which a face is perceived [73]. For this study, we selected the subset of 93 photographs of white males with neutral expression.

Using CFD supposes another advantage for our study. Each photograph is accompanied by information about the target face, and it has been rated by a large sample of participants on several social traits. We selected the following social traits: Afraid, Angry, Attractive, Baby-Faced, Disgusted, Dominant, Feminine, Happy, Masculine, Prototypic, Sad, Surprised, Threatening, Trustworthy, and Unusual. Participants responded on a 1–7 Likert scale (1 = not at all, 7 = extremely) except for Prototypic, that was responded on a 1–5 Likert scale. Prototypic was defined as in which degree the face seems typical; in our case, how much their physical features resemble the typical features of white people. Detailed information on the database generation and characteristics of the participants is available in Ma et al. [72].

We developed an algorithm to automatically process images from the database and to extract individual images of the facial features of each face (Figure 4(b)). Our objective was to extract the internal features (eyebrows, eyes, nose, and mouth) and the jaw contour. Two automatic facial landmark detectors were employed, one for the internal features [74] and another one for the jaw contour [75]. Then, each feature was extracted individually, centred within the image and crop so all images of a given type of feature have the same size and alignment.

Using this procedure, five databases of images of each feature were created. Then, eigenfaces (a holistic approach usually applied on whole faces) are used to characterize each facial feature by its global appearance [76] (Figure 4(c)). This method performs a principal components analysis over an ensemble of images to form a set of basis images. These basis images, known as eigenpictures, can be linearly combined to reconstruct images in the original set. This procedure allows for automatic, robust, fast, and objective characterization of the facial features considering their global appearance while summarizing the central information to characterize them. In this case, each facial feature was characterized using 45 eigenvalues. The same value was chosen for all of them in order to facilitate the subsequent clustering process, bearing in mind that the explained variances were about 85% or higher in all cases.

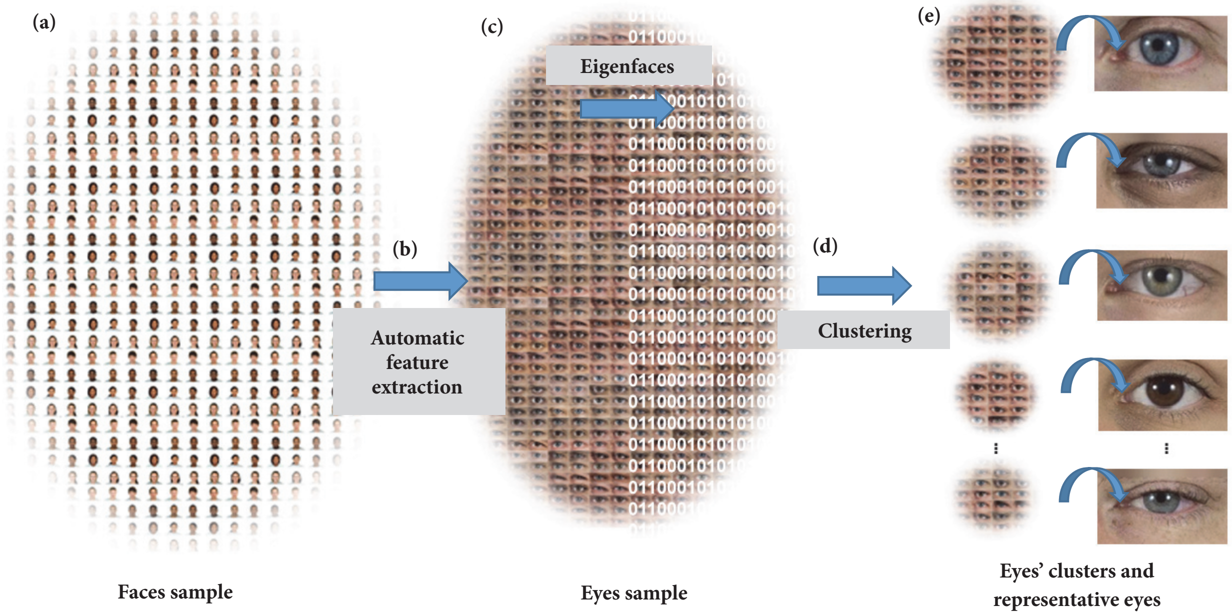

(a)
(b)
Automatic feature extraction
(c)
Eigenfaces
(d)
Clustering
(e)
Faces sample
Eyes sample
Eyes' clusters and representative eyes

Figure 4: Process to establish the alleles of each gene. (a) A set of 93 images of faces is analysed. (b) Individual images of the facial features of each face are automatically extracted. (c) Eigenfaces are used to characterize each facial feature by its global appearance. (d) The facial features are grouped by appearance using their eigenvalues. (e) The features closest to the centre of their clusters will be used as alleles of the corresponding gene in the chromosomes of the faces.

At this stage, the appearance of each feature could be characterized using 45 real values (eigenvalues). K-Means clustering algorithm [77] was selected to cluster the facial features using their eigenvalues as characteristics (Figure 4(d)). A drawback of using this method is that the number of clusters (K) must be predefined. The approach used to face this problem was to perform several K-Means executions varying K and to calculate Dunn's Index [78] for each set of clusters. Dunn's Index measures the compactness and separation of the clusters obtained for each K. A higher Dunn's Index points to a small intracluster variance and a high intercluster distance; namely, the features included in each cluster are more similar among them and more different from the features belonging to other clusters. Therefore, the number of clusters for each feature was selected as the K that maximized Dunn's Index. Using this procedure, eyebrows were classified in 10 clusters (EB1 to EB10), eyes in 19 (E1 to E19), noses in 12 clusters (N1 to N12), mouths in 9 clusters (M1 to M9), and jaws in 11 (J1 to J11). The classification of the facial features for each face in the CFD can be found in the Supplementary Materials of this work (available here). Finally, the features closest to the centre of their clusters were selected as representatives of their groups, and they will be used as alleles of the corresponding gene in the chromosomes of the faces (Figure 4(e)). In this way, all the features in the sample are represented by some allele that has similar appearance. As an example, Figure 5 shows the 9 mouths selected as representatives (alleles). Each allele represents all the mouths in its cluster. The mouths in clusters M3, M5, M6, and M7 are shown in Figure 5.

3.2. Predicting Social Traits from Facial Features. The GA proposed in this work needs an objective function able to measure the fitness of a chromosome with respect to the social traits profile that is looked for. A social traits profile of a face is composed of the scores of the 15 traits selected in the previous section: Afraid, Angry, Attractive, Baby-Faced, Disgusted, Dominant, Feminine, Happy, Masculine, Prototypic, Sad, Surprised, Threatening, Trustworthy, and Unusual. The fitness function for this problem can be formulated as in (1), being $\boldsymbol{T_t^d}$ the desired score for the social trait **t** and $\boldsymbol{T_t}$ the predicted score for the social trait **t** of the chromosome evaluated. While the scores $\boldsymbol{T_t^d}$are known, the values of $\boldsymbol{T_t}$ must be obtained from 15 models, each of them able to predict how human observers would rate the face represented by a chromosome for one of the 15 social traits.

$$F = \sum_{t=1}^{15} \left| T_t^d - T_t \right| \tag{1}$$

Although how the social traits of a face are perceived depends on the whole face, the individual effect of each feature can explain part of the variation within the faces appraisals [79, 80]. A comprehensive discussion on this approach can be found in [81]. From this point of view, some studies have used additive models of the facial attributes appraisals that explain the majority of the feasible explained variance [82, 83], have related individual facial features to perceptions of the targets' personality [84], or have predicted social traits evaluations from facial features with high accuracy [85]. Obviously, using these additive models some unexplained variation remains

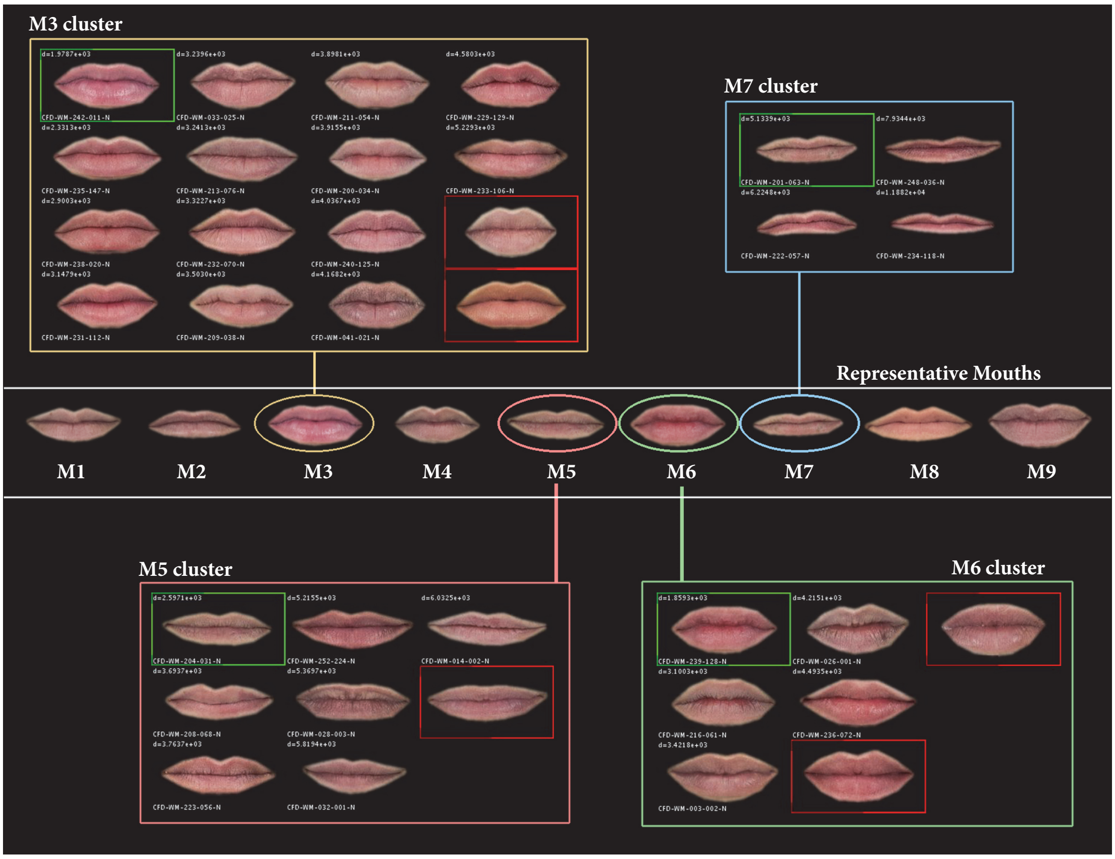


FIGURE 5: Alleles (representatives) M1 to M9 of the mouths. The mouths belonging to clusters M3, M5, M6, and M7 are shown.

due to the interaction among the considered features and because the facial features included in the models do not cover the whole face.

Let us suppose a chromosome with alleles EB, D_{EB}, E, D_E, D_{EE}, N, D_N, M, D_M, and J. To predict $\boldsymbol{Tt}$ (the score of the face represented by this chromosome for the social trait **t**), we propose the additive model shown in (2). In this equation, each $\boldsymbol{S_t^f}$ is the individual score of the allele of the feature **f** assessed with respect to the trait **t**, and $\boldsymbol{w_t^f}$ is the weight of the feature **f** in the assessment of the global face with respect to the trait **t**.

$$T_t = \begin{bmatrix} S_t^{EB} \\ S_t^{D_{EB}} \\ S_t^{E} \\ S_t^{D_E} \\ S_t^{D_{EE}} \\ S_t^{N} \\ S_t^{D_N} \\ S_t^{M} \\ S_t^{D_M} \\ S_t^{J} \end{bmatrix} * \begin{bmatrix} w_t^{EB} \\ w_t^{D_{EB}} \\ w_t^{E} \\ w_t^{D_E} \\ w_t^{D_{EE}} \\ w_t^{N} \\ w_t^{D_N} \\ w_t^{M} \\ w_t^{D_M} \\ w_t^{J} \end{bmatrix}^T \tag{2}$$

The predicted scores of each allele of the feature **f** with respect to each social trait ($\boldsymbol{S_t^f}$) are calculated using (3). In this equation, $\boldsymbol{\overline{S}_t^f}$ is obtained from (4), where **nc** is the number of features in the cluster that is represented by the allele and $\mathbf{S^f_{t\ i}}$ is the score in the social trait **f** of the face to which belongs the cluster member **i**. For example, Figure 6 shows how $\boldsymbol{\overline{S}_t^f}$ is calculated for the M5 allele (of the feature mouth) for a social trait **t**. The mouth M5 (a) is representative of a cluster of mouths (b). Each mouth in this cluster has been extracted from a whole face in the CFD (c), and these faces have scores ($\boldsymbol{S^f_{t\ i}}$) for all the social traits obtained from a group of human observers (d). $\boldsymbol{\overline{S}_t^{M5}}$ is calculated as the mean value of these scores. The scores of each face in the CFD for each social trait can be found in the Supplementary Materials of this work.

$$S_t^f = \frac{\overline{S}_t^f - \mu_{\overline{S}_t^f}}{\sigma_{\overline{S}_t^f}} \bullet \sigma_t^{CFD} + \mu_t^{CFD} \tag{3}$$

$$\overline{S}_t^f = \frac{\sum_{i=1}^{nc} S_{t\ i}^f}{nc} \tag{4}$$

As $\boldsymbol{\overline{S}_t^f}$ are computed using the mean of the scores of the faces of the CFD, the variance of $\boldsymbol{\overline{S}_t^f}$ values is much smaller than that of the scores given by the human raters. So that the

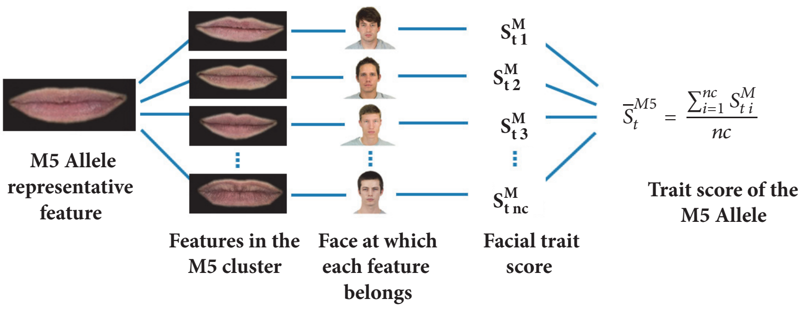


Figure 6: Process to obtain the predicted score of the allele M5.

Table 1: Weight of each feature on social traits appraisals normalized to sum up 1 for each trait.

	Eyebrow	Eye	Nose	Mouth	Jaw	DEB	DE	DN	DM	DEE
Afraid	0.115	0.170	0.085	0.132	0.089	0.053	0.001	0.106	0.139	0.110
Angry	0.108	0.111	0.102	0.046	0.097	0.017	0.159	0.076	0.125	0.159
Attractive	0.183	0.141	0.144	0.104	0.039	0.051	0.048	0.095	0.063	0.133
Baby-faced	0.152	0.120	0.065	0.121	0.096	0.077	0.069	0.058	0.104	0.137
Disgusted	0.113	0.148	0.100	0.079	0.128	0.037	0.032	0.172	0.032	0.159
Dominant	0.085	0.085	0.131	0.036	0.108	0.117	0.161	0.017	0.083	0.176
Feminine	0.087	0.067	0.088	0.046	0.088	0.199	0.041	0.145	0.074	0.166
Happy	0.184	0.171	0.060	0.181	0.095	0.003	0.017	0.068	0.134	0.088
Masculine	0.138	0.097	0.084	0.090	0.111	0.073	0.136	0.053	0.065	0.153
Prototypic	0.168	0.203	0.055	0.163	0.007	0.055	0.094	0.060	0.163	0.033
Sad	0.117	0.115	0.112	0.105	0.059	0.149	0.028	0.000	0.148	0.166
Surprised	0.123	0.119	0.091	0.104	0.094	0.128	0.062	0.085	0.104	0.091
Threat.	0.084	0.148	0.084	0.028	0.123	0.072	0.173	0.036	0.089	0.162
Trust.	0.135	0.184	0.016	0.163	0.092	0.083	0.056	0.141	0.054	0.075
Unusual	0.127	0.117	0.058	0.102	0.129	0.085	0.134	0.068	0.086	0.095
Mean	0.128	0.133	0.085	0.100	0.090	0.080	0.081	0.079	0.098	0.127

models can take extreme values present in the CFD scores, $\overline{\boldsymbol{S}}_{\boldsymbol{t}}^{\boldsymbol{f}}$ are transformed like in (3). In this equation, $\boldsymbol{\mu}_{\overline{S}_t^f}$ and $\boldsymbol{\sigma}_{\overline{S}_t^f}$ are the mean and the standard deviation of the $\overline{\boldsymbol{S}}_{\boldsymbol{t}}^{\boldsymbol{f}}$ values of all the alleles of the feature **f** for the trait **t**, and $\boldsymbol{\mu}_{\boldsymbol{t}}^{\boldsymbol{CFD}}$ and $\boldsymbol{\sigma}_{\boldsymbol{t}}^{\boldsymbol{CFD}}$ are the mean and the standard deviation of the scores in the CFD for the trait **t**. In this way, $\boldsymbol{S}_{\boldsymbol{t}}^{\boldsymbol{f}}$ values have the same mean and standard deviation as the original CFD scores.

The individual effect of each feature can explain part of the variation within the faces appraisals [79, 80], but each facial feature has different effect size. Using a weight per facial feature and social trait, like in (2), gives different importance to each facial feature on the formation of the impression of each social trait. The capability of the developed models to predict the perceived social traits lies in achieving a good fitting to the scores of human observers (available on the CFD). Therefore, it is necessary to find the best combination of weights. To do that, all the faces in the CFD were codified as their corresponding chromosomes. Then, we used a GA in which the fitness function was defined as the mean squared error between the model predictions on the chromosomes and the actual face scores of the assessed faces. Given the characteristics of the problem, using gradient-based methods such as Quasi-Newton method might be sufficient in this case; however, we used a GA because the structure of our big dataset was well conditioned to be used by our calculation module, and using another procedure would have required a time-consuming dataset processing.

The GA was configured to perform single-point crossover and uniform mutation. The crossover probability was set at 0.6 and the mutation probability at 0.001 on a population of 50 individuals. The permitted range for the weights was set to the interval [0; 1]. The selection method employed was Stochastic Universal Sampling, and the Survivor Selection Policy was fitness-based with elitism. The number of iterations was established at 200 000; however, this limit was never reached due to the early stopping condition implemented. This condition allowed for a maximum of 100 consecutive iterations without a change higher than 0.0001 in the best solution fitness. With this configuration, the optimization was performed individually for each social trait, resulting in a total of 15 sets of weights, one for each trait. The obtained weights, normalized to sum up 1 for each social trait, are shown in Table 1. Table 2 shows Pearson's r correlation coefficient and mean square errors (MSE) between the results

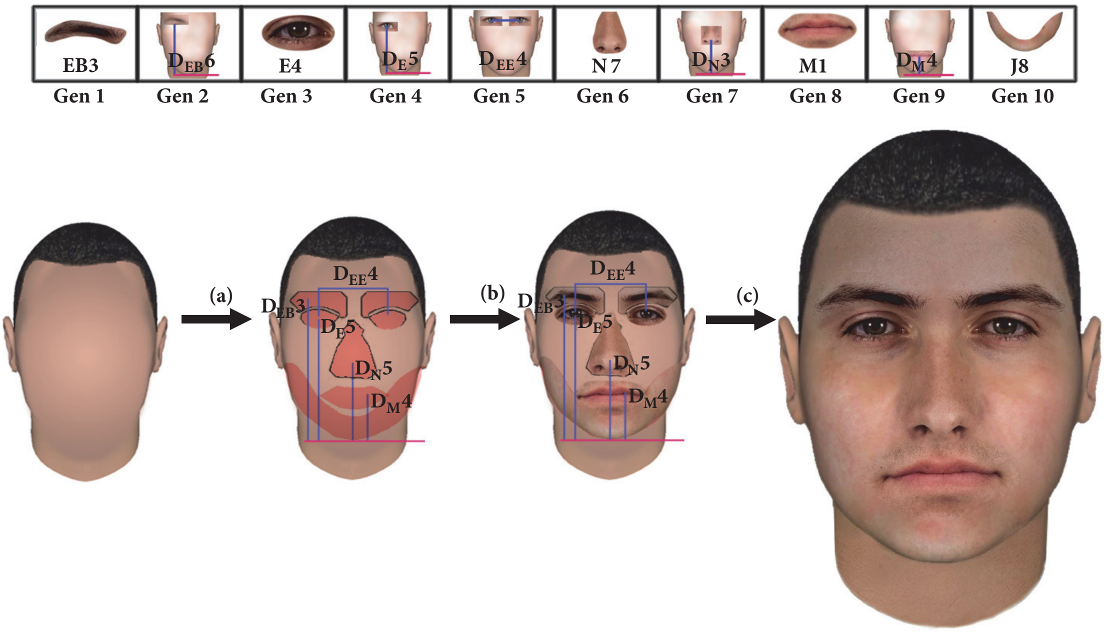


Figure 7: Generating a realistic looking face from a chromosome combining facial features. (a) Mask positioning. (b) Pasting the features. (c) Applying the Poisson Image Editing method.

Table 2: Pearson's r correlation coefficient and mean square errors (MSE) between the results of the models and the actual faces scores by social trait.

Social trait	r	MSE
Afraid	0.70	0.1013
Angry	0.73	0.1872
Attractive	0.77	0.1923
Baby-faced	0.81	0.1661
Disgusted	0.70	0.0841
Dominant	0.74	0.2480
Feminine	0.81	0.0528
Happy	0.76	0.1222
Masculine	0.79	0.1051
Prototypical	0.82	0.7067
Sad	0.73	0.2183
Surprised	0.78	0.0220
Threatening	0.76	0.1730
Trustworthy	0.75	0.0633
Unusual	0.75	0.1896
Mean	0.76	0.1755

of the models and the actual faces scores. All the correlations were highly significant (p values under 0.01).

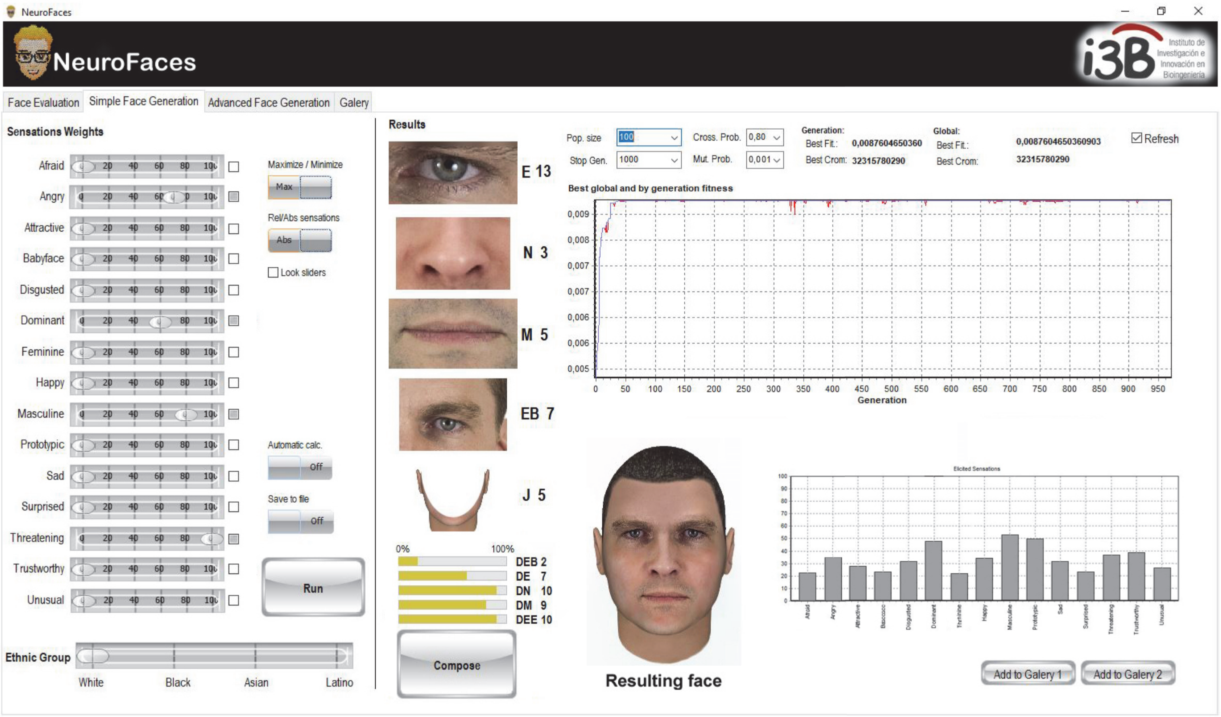
4. Generating Realistic Looking Faces from Chromosomes

Once the GA has found the optimal combination of facial features for eliciting a preestablished social traits profile, it is necessary to generate a realistic looking face combining these facial features. In order to achieve a realistic face, it is necessary to use an automatic seamless fusion method, which further adapts the illumination and tone of the different patches being sewed. The algorithm used in this work to achieve this task is the Poisson Image Editing method [86]. This algorithm makes use of the Poisson Equation and information of the gradient of the images in order to achieve a seamless fusion.

The process is depicted in Figure 7. A base face in which to paste the different features was generated using FaceGen software [87]. This base face is common for all the faces. The genes that codify the facial features (1, 3, 6, 8, and 10) are used to get the images corresponding to the facial features to be pasted and to create masks using the landmarks of the features. The masks are positioned over the base face in the positions established in the genes 2, 4, 5, 7, and 9 of chromosome (Figure 7(a)). Then, the images of each feature are pasted over the corresponding mask (Figure 7(b)). Finally, the Poisson Image Editing method automatically configures the new face.

5. Materials and Methods

A software implementing the GA and the Poisson Image Editing method was developed (Figure 8). This application permits two different tasks. On the one hand, it makes evaluating an existing face obtaining its predicted social traits profile possible. On the other hand, the software allows defining a social traits profile to be obtained, establishing the



Pop. size 100
Cross. Prob. 0,80
Stop Gen. 1000
Mut. Prob. 0,001
Generation:
Best Fit.: 0,0087604650360
Best Crom: 32315780290
Global:
Best Fit.: 0,0087604650360903
Best Crom: 32315780290
☑ Refresh

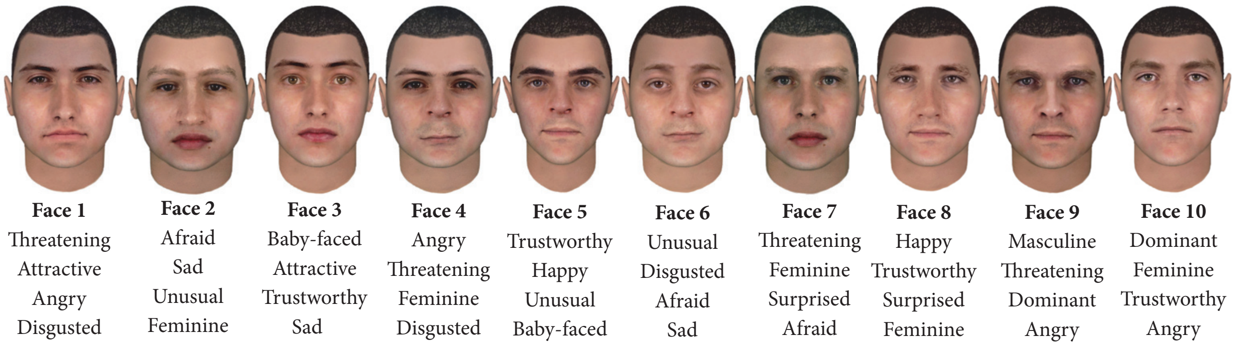


Figure 8: Software implementing the genetic algorithm and the Poisson Image Editing method, and 10 faces generated using the software.

parameters of the GA, and generating a realistic looking face corresponding to the best chromosome found by the GA.

10 faces were generated using the software to test the performance of the GA to produce faces that elicit a preestablished social traits profile and the capacity of the models developed to predict the sensations elicited. To generate the faces, 10 different social traits profiles were used. The software was used to obtain 10 faces from these profiles. The 10 faces are shown in Figure 8. The objective was to compare the social traits profiles of the obtained faces with the opinion of human evaluators.

We must distinguish here between the desired profile of social traits that we initially established as objective and the profiles finally obtained for the faces. There are correlations between the perceived social traits. For example, a highly masculine face is usually perceived as dominant [72] or a baby-faced one as trustworthy. Some of the profiles used to generate the faces combined some usually highly correlated social traits like Masculine, Threatening and Dominant, or Baby-Faced and Trustworthy (like the faces 9 or 3 in Figure 8). In these cases, the algorithm was able to find a combination of facial features with a social traits profile very similar to the desired profile. On the other hand, some other desired profiles joint an unusual combination of social traits, like Dominant and Feminine (face 10), or Angry, Threatening, and Feminine (face 4), or include Unusual as a main social trait (faces 2, 5, and 6). These combinations include social traits that have negative correlations [72]. This means, for example, that changing a facial feature in a given face to increase the perception of Dominant will decrease the perception of Feminine. In these cases, the algorithm will find the face with the social traits profile nearest to the desired one; however, the differences between them will increase as the negative correlation between the desired social traits increases. In some extreme cases, the profile of the face finally obtained could be far of the desired one, for example, if the desired profile includes Feminine and Masculine simultaneously. In these cases, there is no

Table 3: Pearson's r correlation coefficient, p values, and MSE between the predicted scores and the actual faces scores by social trait.

Social trait	r	P value	MSE
Afraid	0.7138	0.0204*	0.3183
Angry	0.4555	0.1859	0.5461
Attractive	0.6635	0.0365*	1.1713
Baby-faced	0.7081	0.0219*	0.4359
Disgusted	0.1993	0.5809	0.7178
Dominant	0.7444	0.0135*	0.4712
Feminine	0.7992	0.0055*	0.1928
Happy	0.2829	0.4284	0.8351
Masculine	0.8222	0.0035*	0.7031
Prototypic	0.1410	0.6977	0.9038
Sad	0.6751	0.0322*	0.2351
Surprised	0.5437	0.1042	1.1461
Threatening	0.5429	0.1048	0.8069
Trustworthy	0.1930	0.5931	0.7370
Unusual	0.6575	0.0388*	0.3852

combination of facial features that can achieve a social traits profile as the desired one.

Under each face in Figure 8, the 4 main social traits we used to define its desired profile are shown.

6. Results and Discussion

This work proposes an evolutionary algorithm to automatically create virtual realistic faces that convey 15 facial social traits, each of them in a predefined quantity, combining the appropriate set of facial features to form the faces. For each social trait, a model that predicts the scores of human raters has been developed. 10 faces with different social traits profiles were generated using the proposed procedure. To test the performance of the system, the results were compared with the opinion of human evaluators. 35 people participated in the survey, 16 men and 19 women. The ages of the participants were between 18 and 71 years old, with a mean age of 37. Participants were asked to assess the 10 created faces using the same scale as the CFD (1–7 Likert). To avoid the learning effect, the social traits and the face order were randomly presented to each participant.

Table 3 shows Pearson's r correlation coefficient, p values, and MSE between the predicted scores and the actual faces scores by social trait. Positive correlations were found for all the traits, being strong and statistically significant for 8 of them, namely, Afraid, Attractive, Baby-Faced, Dominant, Feminine, Masculine, Sad, and Unusual. Low MSE between the predicted scores and the actual faces scores by social trait were obtained for these traits. Although moderate positive correlations were found for Angry, Surprised, and Threatening, these were not significant.

The main objective of this work was to generate faces that elicit a preestablished set of social traits on most observers. Figure 9 shows the results for each face. Blue bars represent the social traits profile predicted by the models. The orange lines are the mean of the scores of human participants (whiskers represent ± 1 times the standard deviation about the mean). The MSE between predicted scores and the means of the scores of the participants are shown for each face in Figure 9. The mean MSE between the predicted scores and the actual faces scores of 10 faces generated by the proposed system was lower than 0.64. Considering only the 8 social traits in which significant correlations were found (Afraid, Attractive, Baby-Faced, Dominant, Feminine, Masculine, Sad, and Unusual), the mean MSE for all the faces was 0.26.

Despite the complexity of the face perception process, the results obtained show that 8 of the models developed in this work have been able to establish the relationships between the facial features and the social traits elicited in the observers. In addition, the interrater agreement among people's judgements on social traits of faces is usually low [72]. However, the proposed procedure was able to approximate the mean opinion of the human observers, finding strong correlations for these 8 social traits.

On the other hand, finding the combination of facial features that elicits several social traits simultaneously, each of them in a predefined amount, is a complex multiobjective problem. This work approached the problem using eigenfaces to create clusters of facial features with the same appearance and selecting one representative feature of each cluster to be used as alleles in a GA. The mean MSE obtained for the tested faces (0.26 on a 1–7 Likert scale) suggests the validity of this approach.

The models obtained in this work to predict social traits from facial features give insights on how important each facial feature is in the formation of each impression of a face. Each additive model considers the individual contribution of each facial feature to explain part of the variation within the appraisals of a social trait. The models add the individual contribution of each feature, weighted by its relative importance in the social trait assessed. The weights presented in Table 1 suggest the effect of each facial feature on the variation of each social trait. For example, in the case of Afraid, the eyes, the mouth, and the position of the mouth seem to have a bigger effect than, for example, the nose or the jaw. Therefore, if it is necessary to change the level in which a given face is perceived as Afraid, shifting the facial features with higher weights will have a bigger effect.

Even though there exists some works on this topic, any of them allows creating realistic faces conveying more than one social trait at a time. Dotsch and Todorov [45] use grey images with superimposed noise in order to achieve faces which convey trustworthiness or dominance. Vernon et al. [88] propose a system able to model social traits and produce cartoon-like computer-generated faces able to elicit three social traits: approachability, youthfulness, and dominance. Perhaps, the proposal closest to the one presented in this work is the one of Walker and Vetter [49], which is capable of creating realistic faces expressing only one social trait at a time. According to our best knowledge, this is the most comprehensive work, in terms of number of social traits considered, generating realistic looking faces that elicit a preestablished set of sensations on most observers.

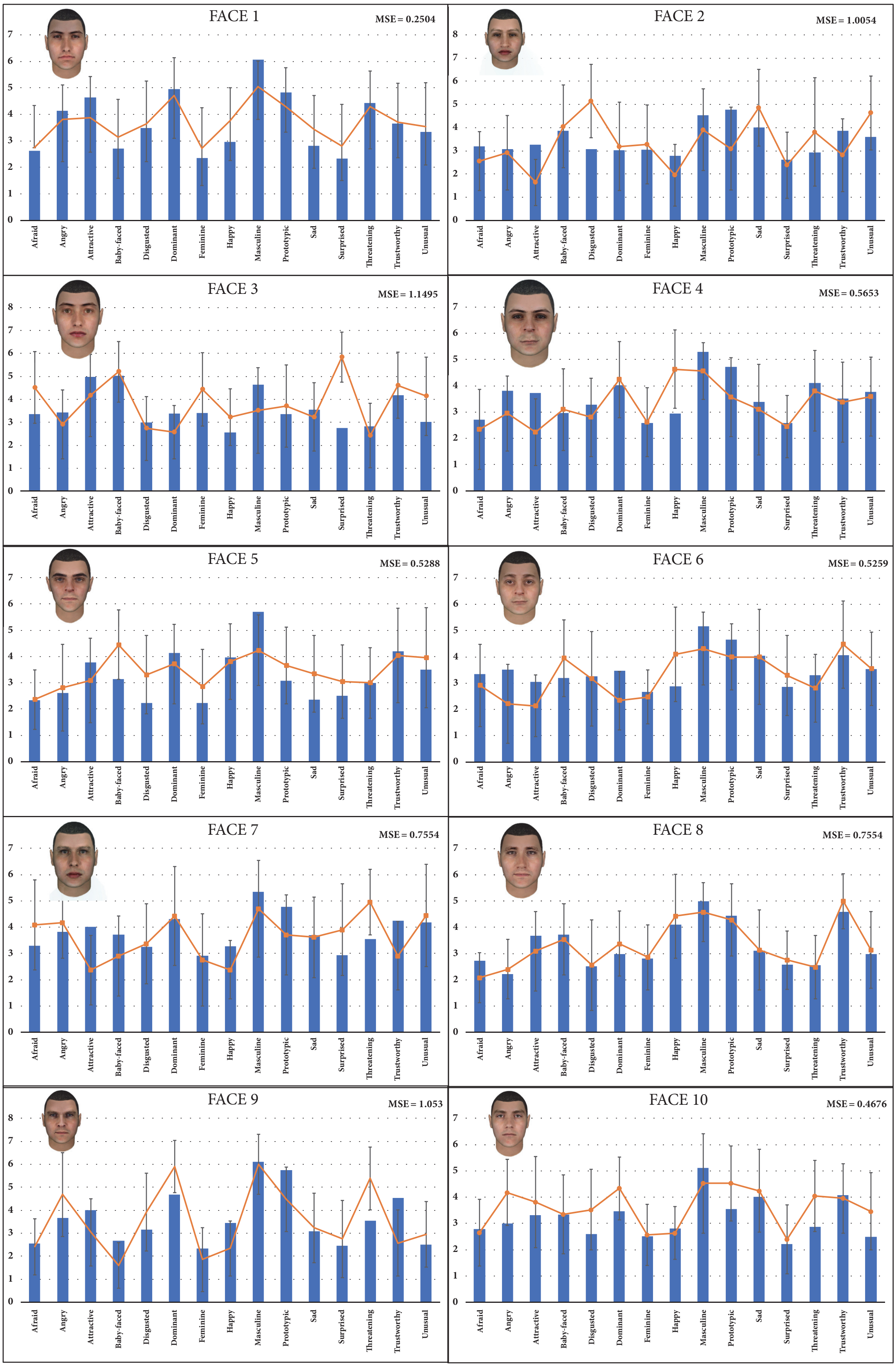


Figure 9: Scores for each face. Blue bars represent the social traits profile predicted by the models. The orange lines are the mean of the scores of human participants (whiskers represent ± 1 times the standard deviation about the mean).

However, some limitations of this study must be pointed out, mainly regarding the generalization of the findings. 93 faces of the Chicago Face Database were used to obtain the models relating facial features and facial assessments. The set of faces belongs to men between the ages of 18 and 40 years living in the Chicago (USA) area. The subjective classifications of the faces were made by a specific group of women and men probably from the same city [72]. Therefore, both the faces and the appraisals used to develop the models come from a specific community. The generalization of the results to faces of people from other communities must be carefully addressed.

Our future works will be intended for developing similar studies for female faces and for extending the results to other races. On the other hand, visual perception research has shown that human brain processes faces in a very complex way [30]. Although the first-order features play a central role in how a face is perceived, second- and higher-order features emerge from the combination of several lower-order features and are used to make judgments from faces. Using a larger face database in our future works would allow us to consider interactions between the facial features, at least of second order, and, probably, to improve the results obtained.

7. Conclusions

This work proposes a new approach to automatically create virtual realistic faces that convey several social traits simultaneously, each of them in a predefined quantity. To create the faces, a genetic algorithm selects the appropriate facial features (including eyes, eyebrows, nose, mouth, and jaw) and their relative positions, in such a way that impressions elicited in observers are as similar as possible to those established by the designer. The facial features used by the algorithm as alleles are obtained using the eigenfaces method. Using this method clusters of facial features with the same appearance were created, and one representative feature of each cluster is used as alleles. Several models that relate the facial features of the faces to the social traits perceived by human observers were developed. These models are used as the fitness function of the genetic algorithm. Finally, the Poisson Image Editing method is used to combine the selected facial features in a face.

15 models were developed to establish the relationships between the facial features and the social traits elicited in human observers. Positive, strong, and statistically significant correlations were found for 8 of them, namely, Afraid, Attractive, Baby-Faced, Dominant, Feminine, Masculine, Sad, and Unusual. To test the proposed procedure, several social traits profiles were established and the developed system was used to generate faces with these social traits. The social traits of the generated faces predicted by the models were compared to the opinion of human observers. The mean squared error obtained for the tested faces (0.26 on a 1–7 Likert scale) suggests the validity of this approach and that the system is able to approximate the mean opinion of the human observers.

Using the developed system, the designer can establish the amount of each social trait that must be elicited by a face, and the system automatically generates the proper face. People use information from faces to judge the emotions and intentions of the owners of the faces. How a face looks in a painting or an advertisement can dramatically influence what we feel about them and what emotions are elicited. In these fields, the procedure presented in this work can be used for creating faces that conveys the desired set of sensations to the observer. In the same way, it can be used in other fields like online activities or new human-machine interaction systems in which it is common to use human digital representations that symbolize the user’s presence or that act as virtual interlocutor.

Data Availability

The Chicago Face Database used to support the findings of this study is freely accessible on http://faculty.chicagobooth.edu/bernd.wittenbrink/cfd/index.html. All the images employed in this study and the results of the facial features clustering are available on https://github.com/flifuehu/.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Supplementary Materials

Scores for each social trait in CFD and classification of the facial features for each face. *(Supplementary Materials)*

References

[1] S. Blick and C. T. Little, “Set in Stone: The Face in Medieval Sculpture,” *The Sixteenth Century Journal*, vol. 39, no. 3, p. 924, 2008.

[2] V. Bruce and A. Young, *Face perception*, Psychology Press, New York, NY, 2012.

[3] A. Todorov, R. Dotsch, D. H. J. Wigboldus, and C. P. Said, “Data-driven methods for modeling social perception,” *Social and Personality Psychology Compass*, vol. 5, no. 10, pp. 775–791, 2011.

[4] A. R. Damasio, “Prosopagnosia,” *Trends in Neurosciences*, vol. 8, no. C, pp. 132–135, 1985.

[5] V. Bruce and A. Young, “Understanding face recognition,” *British Journal of Psychology*, vol. 77, no. 3, pp. 305–327, 1986.

[6] A. C. Little, R. P. Burriss, B. C. Jones, and S. C. Roberts, “Facial appearance affects voting decisions,” *Evolution and Human Behavior*, vol. 28, no. 1, pp. 18–27, 2007.

[7] A. Todorov, S. Fiske, and D. Prentice, *Evaluating Faces on Social Dimensions. Social Neuroscience: Toward Understanding the Underpinnings of the Social Mind*, 2011.

[8] L. A. Zebrowitz and J. M. Montepare, “Social Psychological Face Perception: Why Appearance Matters,” *Social and Personality Psychology Compass*, vol. 2, no. 3, pp. 1497–1517, 2008.

[9] A. Todorov, C. P. Said, A. D. Engell, and N. N. Oosterhof, “Understanding evaluation of faces on social dimensions,” *Trends in Cognitive Sciences*, vol. 12, no. 12, pp. 455–460, 2008.

[10] A. Todorov, A. N. Mandisodza, A. Goren, and C. C. Hall, "Inferences of competence from faces predict election outcomes," *Science*, vol. 308, no. 5728, pp. 1623–1626, 2005.

[11] J. L. Eberhardt, P. G. Davies, V. J. Purdie-Vaughns, and S. L. Johnson, "Looking deathworthy perceived stereotypicality of black defendants predicts capital-sentencing outcomes," *Psychological Science*, vol. 17, no. 5, pp. 383–386, 2006.

[12] J. P. Wilson and N. O. Rule, "Facial Trustworthiness Predicts Extreme Criminal-Sentencing Outcomes," *Psychological Science*, vol. 26, no. 8, pp. 1325–1331, 2015.

[13] J. Bovet, J. Barthes, V. Durand, M. Raymond, A. Alvergne, and A. Sánchez, "Men's Preference for Women's Facial Features: Testing Homogamy and the Paternity Uncertainty Hypothesis," *PLoS ONE*, vol. 7, no. 11, p. e49791, 2012.

[14] B. J. W. Dixson, D. Sulikowski, A. Gouda-Vossos, M. J. Rantala, and R. C. Brooks, "The masculinity paradox: facial masculinity and beardedness interact to determine women's ratings of men's facial attractiveness," *Journal of Evolutionary Biology*, vol. 29, no. 11, pp. 2311–2320, 2016.

[15] C. F. Keating and J. Doyle, "The faces of desirable mates and dates contain mixed social status cues," *Journal of Experimental Social Psychology*, vol. 38, no. 4, pp. 414–424, 2002.

[16] J. H. Langlois, L. Kalakanis, A. J. Rubenstein, A. Larson, M. Hallam, and M. Smoot, "Maxims or myths of beauty? A meta-analytic and theoretical review," *Psychological Bulletin*, vol. 126, no. 3, pp. 390–414, 2000.

[17] E. Liaci, A. Fischer, M. Heinrichs, L. T. van Elst, and J. Kornmeier, "Mona Lisa is always happy – and only sometimes sad," *Scientific Reports*, vol. 7, no. 1, 2017.

[18] M. Cerf, E. P. Frady, and C. Koch, "Faces and text attract gaze independent of the task: experimental data and computer model," *Journal of Vision*, vol. 9, no. 12, pp. 74–76, 2009.

[19] R. Jack and P. Schyns, "The Human Face as a Dynamic Tool for Social Communication," *Current Biology*, vol. 25, no. 14, pp. R621–R634, 2015.

[20] A. Davis, J. Murphy, D. Owens, D. Khazanchi, and I. Zigurs, "Avatars, people, and virtual worlds: Foundations for research in metaverses," *Journal of the Association for Information Systems*, vol. 10, no. 2, pp. 90–117, 2009.

[21] N. Yee and J. Bailenson, "The proteus effect: the effect of transformed self-representation on behavior," *Human Communication Research*, vol. 33, no. 3, pp. 271–290, 2007.

[22] M. Fabri and D. Moore, "The use of emotionally expressive avatars in Collaborative Virtual Environments," in *Proceedings of the AISB'05 Convention: Social Intelligence and Interaction in Animals, Robots and Agents - Joint Symposium on Virtual Social Agents: Social Presence Cues for Virtual Humanoids Empathic Interaction with Synthetic Characters Mind Minding Agents*, pp. 88–94, UK, April 2005.

[23] M. Fabri, S. Elzouki, and D. Moore, "Emotionally expressive avatars for chatting, learning and therapeutic intervention," in *Human-Computer Interact*, J. A. Jacko, Ed., pp. 275–285, Springer, Berlin, Heidelberg, Germany, 2007.

[24] P. V. R. Carvalho, I. L. dos Santos, J. O. Gomes, M. R. S. Borges, and S. Guerlain, "Human factors approach for evaluation and redesign of human-system interfaces of a nuclear power plant simulator," *Displays*, vol. 29, no. 3, pp. 273–284, 2008.

[25] V. Orvalho, J. Miranda, and A. A. Sousa, "Facial Synthesys of 3D Avatars for Therapeutic Applications," *Studies in Health Technology and Informatics*, vol. 144, pp. 96–98, 2009.

[26] T. Trescak, A. Bogdanovych, S. Simoff, and I. Rodriguez, "Generating diverse ethnic groups with genetic algorithms," in *Proceedings of the 18th ACM symposium on Virtual reality software and technology - VRST '12*, ACM Press, NY, USA, 2012.

[27] J. A. Diego-Mas and J. Alcaide-Marzal, "A computer based system to design expressive avatars," *Computers in Human Behavior*, vol. 44, pp. 1–11, 2015.

[28] A. Albin-Clark and T. Howard, *Automatically Generating Virtual Humans using Evolutionary Algorithms*, W. Tang and J. Collomosse, Eds., EG UK Theory and Practice of Computer Graphics, 2009.

[29] P. Sukhija, S. Behal, and P. Singh, "Face Recognition System Using Genetic Algorithm," *Procedia Computer Science*, vol. 85, pp. 410–417, 2016.

[30] D. W. Piepers and R. A. Robbins, "A Review and Clarification of the Terms "holistic," "configural," and "relational" in the Face Perception Literature," *Frontiers in Psychology*, vol. 3, 2012.

[31] I. Biederman, "Recognition-by-components: a theory of human image understanding," *Psychological Review*, vol. 94, no. 2, pp. 115–147, 1987.

[32] R. Diamond and S. Carey, "Why Faces Are and Are Not Special. An Effect of Expertise," *Journal of Experimental Psychology: General*, vol. 115, no. 2, pp. 107–117, 1986.

[33] J. C. Bartlett, J. H. Searcy, and H. Abdi, *What Are the Routes to Face Recognition? Perception of Faces, Objects, and Scenes: Analytic and Holistic Processes*, 2006.

[34] V. Goffaux and B. Rossion, "Faces are "spatial"–holistic face perception is supported by low spatial frequencies.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 32, no. 4, pp. 1023–1039, 2006.

[35] E. McKone, "Isolating the Special Component of Face Recognition: Peripheral Identification and a Mooney Face.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 30, no. 1, pp. 181–197, 2004.

[36] J. W. Tanaka and M. J. Farah, "Parts and Wholes in Face Recognition," *The Quarterly Journal of Experimental Psychology Section A*, vol. 46, no. 2, pp. 225–245, 2018.

[37] B. Rossion, "Picture-plane inversion leads to qualitative changes of face perception," *Acta Psychologica*, vol. 128, no. 2, pp. 274–289, 2008.

[38] E. Mckone and G. Yovel, "Why does picture-plane inversion sometimes dissociate perception of features and spacing in faces, and sometimes not? toward a new theory of holistic processing," *Psychonomic Bulletin & Review*, vol. 16, no. 5, pp. 778–797, 2009.

[39] H. R. Wilson, G. Loffler, and F. Wilkinson, "Synthetic faces, face cubes, and the geometry of face space," *Vision Research*, vol. 42, no. 27, pp. 2909–2923, 2002.

[40] N. Thalmann, P. Kalra, and M. Escher, "Face to virtual face," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 870–883, 1998.

[41] K. E. Ketchum, "Facegen and the Technovisual Politics of Embodied Surfaces," *WSQ: Women's Studies Quarterly*, vol. 37, no. 1-2, pp. 183–199, 2009.

[42] E. B. Roesch, L. Tamarit, L. Reveret, D. Grandjean, D. Sander, and K. R. Scherer, "FACSGen: A Tool to Synthesize Emotional Facial Expressions Through Systematic Manipulation of Facial Action Units," *Journal of Nonverbal Behavior*, vol. 35, no. 1, pp. 1–16, 2011.

[43] M. C. Mangini and I. Biederman, "Making the ineffable explicit: Estimating the information employed for face classifications," *Cognitive Science*, vol. 28, no. 2, pp. 209–226, 2004.

[44] R. Dotsch, D. H. J. Wigboldus, O. Langner, and A. Van Knippenberg, “Ethnic out-group faces are biased in the prejudiced mind,” *Psychological Science*, vol. 19, no. 10, pp. 978–980, 2008.

[45] R. Dotsch and A. Todorov, “Reverse Correlating Social Face Perception,” *Social Psychological and Personality Science*, vol. 3, no. 5, pp. 562–571, 2012.

[46] A. J. Calder and A. W. Young, “Understanding the recognition of facial identity and facial expression,” *Nature Reviews Neuroscience*, vol. 6, no. 8, pp. 641–651, 2005.

[47] N. N. Oosterhof and A. Todorov, “The functional basis of face evaluation,” *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 105, no. 32, pp. 11087–11092, 2008.

[48] C. P. Said, N. Sebe, and A. Todorov, “Structural Resemblance to Emotional Expressions Predicts Evaluation of Emotionally Neutral Faces,” *Emotion*, vol. 9, no. 2, pp. 260–264, 2009.

[49] M. Walker and T. Vetter, “Portraits made to measure: Manipulating social judgments about individuals with a statistical face model,” *Journal of Vision*, vol. 9, no. 11, pp. 12-12, 2009.

[50] S. J. Gibson, C. J. Solomon, M. I. S. Maylin, and C. Clark, “New methodology in facial composite construction: From theory to practice,” *International Journal of Electronic Security and Digital Forensics*, vol. 2, no. 2, pp. 156–168, 2009.

[51] C. D. Frowd, P. J. B. Hancock, and D. Carson, “EvoFIT: A Holistic, Evolutionary Facial Imaging Technique for Creating Composites,” *ACM Transactions on Applied Perception*, vol. 1, no. 1, pp. 19–39, 2004.

[52] M. S. Keil and K. J. Friston, ““I Look in Your Eyes, Honey”: Internal Face Features Induce Spatial Frequency Preference for Human Face Processing,” *PLoS Computational Biology*, vol. 5, no. 3, p. e1000329, 2009.

[53] D. G. Kwart, T. Foulsham, and A. Kingstone, “Age and beauty are in the eye of the beholder,” *Perception*, vol. 41, no. 8, pp. 925–938, 2012.

[54] E. Fox and L. Damjanovic, “The eyes are sufficient to produce a threat superiority effect,” *Emotion*, vol. 6, no. 3, pp. 534–539, 2006.

[55] C. Saavedra, P. Smith, and J. Peissig, “The Relative Role of Eyes, Eyebrows, and Eye Region in Face Recognition,” *Journal of Vision*, vol. 13, no. 9, pp. 410-410, 2013.

[56] J. Sadr, I. Jarudi, and P. Sinha, “The role of eyebrows in face recognition,” *Perception*, vol. 32, no. 3, pp. 285–293, 2003.

[57] D. Lundqvist, F. Esteves, and A. Öhman, “The face of wrath: Critical features for conveying facial threat,” *Cognition & Emotion*, vol. 13, no. 6, pp. 691–711, 1999.

[58] C. Blais, C. Roy, D. Fiset, M. Arguin, and F. Gosselin, “The eyes are not the window to basic emotions,” *Neuropsychologia*, vol. 50, no. 12, pp. 2830–2838, 2012.

[59] R. L. Terry, “Further Evidence on Components of Facial Attractiveness,” *Perceptual and Motor Skills*, vol. 45, no. 1, pp. 130-130, 2016.

[60] V. Axelrod and G. Yovel, “External facial features modify the representation of internal facial features in the fusiform face area,” *NeuroImage*, vol. 52, no. 2, pp. 720–725, 2010.

[61] A. J. Logan, G. E. Gordon, and G. Loffler, “Contributions of individual face features to face discrimination,” *Vision Research*, vol. 137, pp. 29–39, 2017.

[62] M. K. Yamaguchi, T. Hirukawa, and S. Kanazawa, “Judgment of gender through facial parts,” *Perception*, vol. 42, no. 11, pp. 1253–1265, 2013.

[63] P. M. Pallett, S. Link, and K. Lee, “New “golden” ratios for facial beauty,” *Vision Research*, vol. 50, no. 2, pp. 149–154, 2010.

[64] B. C. Jones, A. C. Little, D. M. Burt, and D. I. Perrett, “When facial attractiveness is only skin deep,” *Perception*, vol. 33, no. 5, pp. 569–576, 2004.

[65] N. Hagiwara, D. A. Kashy, and J. Cesario, “The independent effects of skin tone and facial features on Whites’ affective reactions to Blacks,” *Journal of Experimental Social Psychology*, vol. 48, no. 4, pp. 892–898, 2012.

[66] E. Tsankova and A. Kappas, “Facial Skin Smoothness as an Indicator of Perceived Trustworthiness and Related Traits,” *Perception*, vol. 45, no. 4, pp. 400–408, 2015.

[67] B. Fink, N. Neave, J. T. Manning, and K. Grammer, “Facial symmetry and judgements of attractiveness, health and personality,” *Personality and Individual Differences*, vol. 41, no. 3, pp. 491–499, 2006.

[68] D. E. Goldberg, *Genetic algorithms in search, optimization and machine learning*, Addison-Wesley Longman Publishing Co., Inc, Boston, MA, USA, 1989.

[69] J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, University of Michigan Press, Oxford, UK, 1975.

[70] M. Srinivas and L. M. Patnaik, “Genetic algorithms: a survey,” *The Computer Journal*, vol. 27, no. 6, pp. 17–26, 1994.

[71] M. Chihaoui, A. Elkefi, W. Bellil, and C. Ben Amar, “A Survey of 2D Face Recognition Techniques,” *The Computer Journal* , vol. 5, no. 4, p. 21, 2016.

[72] D. S. Ma, J. Correll, and B. Wittenbrink, “The Chicago face database: A free stimulus set of faces and norming data,” *Behavior Research Methods*, vol. 47, no. 4, pp. 1122–1135, 2015.

[73] R. Russell, “Sex, beauty, and the relative luminance of facial features,” *Perception*, vol. 32, no. 9, pp. 1093–1107, 2003.

[74] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, “Incremental face alignment in the wild,” in *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR ’14)*, pp. 1859–1866, IEEE, Columbus, Ohio, USA, June 2014.

[75] P. B. Thomas, T. Baltrušaitis, P. Robinson, and A. J. Vivian, “The Cambridge Face Tracker: Accurate, Low Cost Measurement of Head Posture Using Computer Vision and Face Recognition Software,” *Translational Vision Science & Technology*, vol. 5, no. 5, 2016.

[76] L. Sirovich and M. Kirby, “Low-dimensional procedure for the characterization of human faces,” *Journal of the Optical Society of America*, vol. 4, no. 3, pp. 519–524, 1987.

[77] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, p. 14, University of California Press, Berkeley, Calif, USA, 1967.

[78] J. C. Dunn, “Well-separated clusters and optimal fuzzy partitions,” *Journal of Cybernetics*, vol. 4, no. 1, pp. 95–104, 1974.

[79] R. Cabeza and T. Kato, “Features are Also Important: Contributions of Featural and Configural Processing to Face Recognition,” *Psychological Science*, vol. 11, no. 5, pp. 429–433, 2000.

[80] S. S. Rakover, “Featural vs. Configurational information in faces: A conceptual and empirical analysis,” *British Journal of Psychology*, vol. 93, no. 1, pp. 1–30, 2002.

[81] L. Z. McArthur and R. M. Baron, “Toward an ecological theory of social perception,” *Psychological Review*, vol. 90, no. 3, pp. 215–238, 1983.

[82] D. Gill, "Women and men integrate facial information differently in appraising the beauty of a face," *Evolution and Human Behavior*, vol. 38, no. 6, pp. 756–760, 2017.

[83] L. T. Maloney and M. F. Dal Martello, "Kin recognition and the perceived facial similarity of children," *Journal of Vision*, vol. 6, no. 10, article no. 4, pp. 1047–1056, 2006.

[84] S. V. Paunonen, K. Ewan, J. Earthy, S. Lefave, and H. Goldberg, "Facial features as personality cues," *Journal of Personality*, vol. 67, no. 3, pp. 555–583, 1999.

[85] M. Rojas Q., D. Masip, A. Todorov, J. Vitria, and C. I. Baker, "Automatic Prediction of Facial Trait Judgments: Appearance vs. Structural Models," *PLoS ONE*, vol. 6, no. 8, p. e23323, 2011.

[86] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," *ACM Transactions on Graphics*, vol. 22, no. 3, pp. 313–318, 2003.

[87] Singular Inversions, "Facegen modeller," Toronto, ON; 2008.

[88] R. J. Vernon, C. A. Sutherland, A. W. Young, and T. Hartley, "Modeling first impressions from highly variable facial images," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 111, no. 32, pp. E3353–E3361, 2014.