Acta Psychologica 160 (2015) 43–57

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/ locate/actpsy

Computerized measures of visual complexity☆

Penousal Machado [a,*], Juan Romero [b], Marcos Nadal [c], Antonino Santos [b], João Correia [a], Adrián Carballal [b]

[a] CISUC, Department of Informatics Engineering, University of Coimbra, Portugal
[b] Faculty of Computer Science, University of A Coruña, Spain
[c] Department of Basic Psychological Research and Research Methods, University of Vienna, Austria

ARTICLE INFO

Article history:
Received 20 January 2015
Received in revised form 21 May 2015
Accepted 18 June 2015
Available online 10 July 2015

Keywords:
Visual complexity
Psychological aesthetics
Vision
Machine learning

ABSTRACT

Visual complexity influences people's perception of, preference for, and behaviour toward many classes of objects, from artworks to web pages. The ability to predict people's impression of the complexity of different kinds of visual stimuli holds, therefore, great potential for many domains, basic and applied. Here we use edge detection operations and several image metrics based on image compression error and Zipf's law to estimate the visual complexity of images. The experiments involved 800 images, each previously rated by thirty participants on perceived complexity. In a first set of experiments we analysed the correlation of individual features with the average human response, obtaining correlations up to $r_s = .771$. In a second set of experiments we employed Machine Learning techniques to predict the average visual complexity score attributed by humans to each stimuli. The best configurations obtained a correlation of $r_s = .832$. The average prediction error of the Machine Learning system over the set of all stimuli was .096 in a normalized 0 to 1 interval, showing that it is possible to predict, with high accuracy human responses. Overall, edge density and compression error were the strongest predictors of human complexity ratings.

1. Introduction

People's preferences for visual objects, scenes, and displays are the result of various cognitive and affective processes (Chatterjee, 2004; Leder, Belke, Oeberst, & Augustin, 2004). Research has shown that several perceptual features—such as colour, colour combinations, contour, or symmetry—influence people's visual preferences and affective responses (Bertamini, Palumbo, Gheorghes, & Galatsidas, in press; Palmer, Schloss, & Sammartino, 2013; Pecchinenda, Bertamini, Makin, & Ruta, 2014). One of such features, complexity, is believed to have a strong impact on preference and affect, given its relation to arousal (Berlyne, 1971; Marin & Leder, 2013), and has therefore been awarded central roles in psychological models of aesthetic appreciation (Berlyne, 1971; Fechner, 1876). From a basic science perspective, thus, research on how the perceptual features that contribute to visual complexity are processed, and how this processing leads to liking and other affective responses, increases our understanding of one of our species distinctive traits: the capacity for aesthetic appreciation. From an applied perspective, it has implications for the design of architectural spaces (Heath, Smith, & Lim, 2000; Imamoglu, 2000), advertisements (Pieters, Wedel, & Batra, 2010), packages (Reimann, Zaichkowsky, Neuhaus, Bender, & Weber, 2010), web pages (Bauerly & Liu, 2008; Krishen, Kamra, & Mac, 2008; Lavie & Tractinsky, 2004; Moshagen & Thielsch, 2010), and in-vehicle navigation devices (Lavie, Oron-Gilad, & Meyer, 2011), among other domains, where visual complexity impacts both liking and usability.

It has long been believed that two aspects of complexity, order and variety, determine beauty. From this perspective, beauty emerges from "unity in variety" (Tatarkiewicz, 1972). The importance of two different—and sometimes opposing—forces was introduced into experimental psychology by Fechner (1876), who formulated the "principle of unitary connection of the manifold" (Cupchik, 1986), that argued that stimuli are pleasing when they adequately balance complexity and order. Birkhoff (1932) formulated this relation between order and complexity in mathematical terms, and argued that beauty increased with order and decreased with complexity. He defined order on the basis of repetition and redundancy, and complexity as an expression of numerousness. Eysenck's (1941, 1942) studies on the correlation between the aesthetic measure predicted by Birkhoff's (1932) formula and participants' beauty ratings suggested that both order and complexity contribute positively to the appreciation of beauty.

Berlyne (1970, 1971) was probably the first to provide a proper psychobiological explanation for the effects of complexity on preference. Berlyne (1971) posited that the hedonic state resulting from the interaction of reward and aversion brain systems would lead people to prefer intermediate levels of complexity, which was defined according to such

☆ Penousal Machado was supported by the Portuguese Foundation for Science and Technology in the scope of project SBIRC (PTDC/EIA–EIA/115667/2009). Marcos Nadal was supported by research grant Research Grant FFI2010-20759 awarded by the Spanish Ministerio de Ciencia e Innovación.

* Corresponding author at: Departamento de Engenharia Informática da Faculdade de Ciências e Tecnologia da Universidade de Coimbra, Pólo II, Pinhal de Marrocos, 3030-290 Coimbra, Portugal.
E-mail address: machado@dei.uc.pt (P. Machado).

http://dx.doi.org/10.1016/j.actpsy.2015.06.005
0001-6918/

aspects as pattern regularity, amount of elements, their heterogeneity, or the irregularity of the forms (Berlyne, 1963, 1970, 1971; Berlyne, Ogilvie, & Parham, 1968). In Berlyne's framework, order is not orthogonal to complexity, given that disorganization is regarded as a kind of complexity, together with the amount of elements. Several studies were conducted to test this hypothesis, employing diverse visual stimuli. Recent research has shown that their results were strongly conditioned by the way complexity had been defined, manipulated and measured (Nadal, Munar, Marty, & Cela-Conde, 2010).

2. Measuring complexity

It has been known for some time now that people's perception of complexity is not merely a direct reflection of the complexity inherent to visual stimuli. Attneave (1957) noted that "the amount of information contained in a stimulus (from the experimenter's point of view) may vary greatly without changing the apparent complexity of the stimulus" (Attneave, 1957, p. 225). Perception is a constructive process. Although it is based on sensory information, its purpose is not to render the world as it is, but to provide us with an image that we can understand and is coherent with our prior knowledge about the world. In order to do so, perception is guided by inference, hypotheses, and other top-down processes, as well as context, which can strongly influence the appearance of an object. Gestalt psychologists characterized several perceptual processes whereby visual features are joined, segregated and grouped to construct meaningful images, and these processes have a crucial role in determining perceived complexity (Strother & Kubovy, 2003).

Even Berlyne (1974) emphasized "The collative variables [including complexity] are actually subjective, in the sense that they depend on the relations between physical and statistical properties of stimulus objects and processes within the organism. A pattern can be more novel, complex, or ambiguous for one person than for another or, for the same person, at one time than at another". "Nevertheless — he added — many experiments, using rating scales and other techniques, have confirmed that collative properties and subjective informational variables tend, as one would expect, to vary concomitantly with the corresponding objective measures of classical information theory" (Berlyne, 1974, p19).

In principle, thus, it should be possible to arrive at a computational measure of visual complexity. This constitutes an interesting objective for at least two reasons. In a basic sense, measures of images' intrinsic complexity would enable determining the perceptual, cognitive or contextual features that influence perceived complexity, moving closer or away from the objective (computational) measure. In an applied sense, it would allow researchers, designers and engineers to anticipate participants', consumers' and users' aesthetic and affective responses to the complexity in their products, ranging from web pages to architectural facades, and including visual displays of all sorts. This would greatly save the time and costs related with post-production tests and surveys.

One of the most popular way of determining visual complexity has been to derive a set of normative scores by asking large samples of participants to rate sets of stimuli on a number of scales, including complexity (Alario & Ferrand, 1999; Bonin, Peereman, Malardier, Méot, & Chalard, 2003; Snodgrass, 1997). This method, however, has a number of drawbacks. First, people's rating of complexity can be confounded by familiarity (Forsythe, Mulhern, & Sawey, 2008) and style (Nadal et al., 2010). Second, it is only useful for images that have already been produced, and does not allow the prediction of the perceived complexity of images whose production is being planned or under development. In this sense, algorithms represent more fruitful and practical avenue possibility.

In their study on icon abstractness García, Badre, and Stasko (1994) developed an algorithmic measure of complexity. This measure took into account the amount of horizontal, vertical, and diagonal lines, as well as the number of open and closed figures, and letters in each icon. McDougall, Curry, and de Bruijn (1999) used the same measure to quantify the complexity of a new set of figures, and they showed that it correlated well with people's judgement of visual complexity (McDougall, de Bruijn, & Curry, 2000). Given how time consuming it was to calculate this metric, Forsythe, Sheehy, and Sawey (2003) devised an automated system to measure icon complexity. They based this metric on perimeter detection measures and a structural variability measure. Their results showed strong correlations between their metric and the scores provided by García et al. (1994) and McDougall et al.'s (1999) studies, revealing that it is possible to approximate human appraisals of complexity with computational metrics of structural properties of images.

The main drawback of this kind of metrics is its limited application to relatively simple and isolated icons and symbols. Algorithmic measures of complexity for richer stimuli, like pictures from nature, chart displays and art, have tended to be based on algorithmic information theory (Donderi, 2006). In short, this theory postulates that the minimum length of the code required to describe a visual image constitutes a good measure of its complexity (Leeuwenberg, 1969; Simon, 1972). Donderi (2003) showed that compressed file size was a good approximation to this minimum length. Furthermore, JPEG and ZIP compressed file lengths significantly correlated with subjectively rated complexity and predicted search time and errors in tasks involving chart displays (Donderi & McFadden, 2005).

Computational measures have also been applied to attempt to quantify the complexity of artworks. Forsythe, Nadal, Sheehy, Cela-Conde, and Sawey (2011) examined the correlation between people's judgement of complexity for 800 artistic and nonartistic, abstract and representational, visual stimuli and JPEG and GIF compression measures, as well as with a perimeter detection measure. Their results showed that the three computational measures significantly correlated with judged complexity, with GIF compression exhibiting the strongest relation ($r_s = .74$) and perimeter detection the weakest ($r_s = .58$), though there were certain differences according to the kind of stimuli.

Marin and Leder (2013) also compared the extent to which several computational measures correlated with participants' complexity ratings of different kinds of materials. For a subset of stimuli from the International Affective Picture System (Lang, Bradley, & Cuthbert, 2005), they found that TIFF file size ($r_s = .53$) and JPEG file size ($r_s = .52$) correlated strongest with subjective complexity ratings. Similarly to Forsythe et al.'s (2011) work, Marin and Leder (2013) reported that measures of perimeter detection showed weaker correlations ($r_s \sim .44$). For this set of stimuli, the highest correlations were obtained with an edge detection measure: the root mean square contrast (RMS), related to the presence of high-contrast features. In this case, the correlation between complexity ratings and the images' mean contrast values of the RMS contrast map was $r_s = .59$. Interestingly, these results were not mirrored in Marin and Leder's (2013) second experiment, which aimed to examine the relation between the same measures and human complexity ratings of 96 representational paintings. For this set, none of the compressed file size measures correlated significantly with the ratings. In fact, the only measure to correlate significantly with complexity ratings was the standard deviation of the mean values of edge detection based on phase congruency ($r_s \sim .38$).

The discrepancies between Forsythe et al.'s (2011) and Marin and Leder's (2013) results probably have to do with the selected materials and procedure. Whereas Forsythe et al. (2011) excluded affectively moving images, Marin and Leder (2013) selected the images in the two aforementioned experiments to accomplish a balanced variation along the arousal and pleasantness dimensions. The images used by Forsythe et al. (2011) were selected on the basis of pilot experiments to cover a broad range of visual complexity, understood in a general sense as the degree of intricacy; the ones used by Marin and Leder (2013) were either figure-ground compositions or complex visual scenes.

Taking a different approach, Taylor, Micholich, and Jonas (1999) argued that Jackson Pollock's renowned drip paintings were fractal patterns, and that the fractal dimension of such patterns could be quantified. Subsequently, Taylor, Micholich, and Jonas (2002) showed that over a decade the fractal dimension of Pollock's paintings increased almost linearly. Fractal dimension is a measure of how much space is filled by a fractal, and could be understood to reflect some form of visual complexity. Spehar, Clifford, Newell, and Taylor (2003) showed that the fractal dimension of Pollock's art corresponds to the range of maximum preference for fractals in natural images and simulated coastlines. Jones-Smith and Mathur (2006), however, have questioned this use of fractal dimension in Pollock's artworks.

To the best of our knowledge, the work of Machado and Cardoso (1998) constitutes the first attempt to use image compression to estimate the human perception of complexity and of aesthetic value. In this early work they resort to JPEG and Fractal image compression to judge the aesthetic value of images. To assess the approach, the authors submit their system to the Design Judgment Test (Graves, 1948) — a test designed to determine how humans respond to several principles of aesthetic order. The percentage of correct answers obtained by the system depends on its parameterization, ranging from 54.4% to 73.3%, with an average of 64.9% over the considered parametric interval. Eysenck and Castle (1971) report average results for art and non-art students of 64.4% and 60%, with variances below 4%. In later studies similar approaches were used to generate images of arguable aesthetic merit (Machado & Cardoso, 2002; Machado, Romero, & Manaris, 2007; Machado, Romero, Santos, Cardoso, & Pazos, 2007). Machado, Romero, and Manaris (2007) and Machado, Romero, Santos, et al. (2007) used an Artificial Neural Network (Rosenblatt, 1958; Rumelhart et al., 1986) in conjunction with a subset of the features proposed in this paper, obtaining an average success rate of 71.67% in the Design Judgment Test, while in Machado, Romero, and Manaris (2007) and Machado, Romero, Santos, et al. (2007) it was used to identify the author of paintings obtaining, in this case, success rates above 90% for all considered painters. Although the direct comparison of the results of these systems to those of humans is tempting, Machado & Cardoso, (1998), Machado, Romero, and Manaris (2007) and Machado, Romero, Santos, et al. (2007) refrain from making this comparison, warning that it can be misleading. Nevertheless, it is reasonable to state that these results demonstrate the viability of the approach and are competitive with the ones obtained by humans.

3. Automated measures of complexity

In this section we will present the basic assumptions underlying the automated measures of image complexity used in this study. We will first refer to the importance of edges in image perception, then we will analyse the relation between compression and complexity, and finally we will present entropy estimates, such as fractal dimension and Zipf's law metrics.

3.1. Edge detection

Edges in an image usually indicate changes in depth, orientation, illumination, material, object boundaries, and so on. As such, edge detection is vital for human and computer vision (e.g., Palmer, 1999). It is thus reasonable to expect that perceived complexity would relate with two different edge parameters: their quantity and their distribution across the image. Accordingly, images regarded as complex are expected to tend to have (i) a greater number and (ii) a less predictable distribution of edges across the image than simpler images.

Following this line of reasoning, we applied edge detection algorithms in the experiments reported in this paper. This involves transforming the image into a new one where the edges are represented in white and everything else is represented in black (Fig. 1). The number of pixels that represent edges can be estimated by considering the average colour of the image resulting from the edge detection step. We will also estimate the regularity of edge distribution, as described below, by applying image compression procedures to the results of edge detection, and by calculating the ZIPF's law coefficients of the edge-images.

3.2. Complexity and compression

In the scope of Algorithmic Information Theory (AIT), complexity and compression are intimately related concepts (Salomon, 1997). Informally, simple images have redundant information and predictable data — which can be explored to represent them compactly — and are, therefore, compressible. In highly complex images, the value of a pixel cannot be predicted from the remaining ones, no redundancy exists and, therefore, is incompressible. In other words, a simple object can be represented compactly while a complex one requires a lengthy description.

The notions above were expressed formally by Andrey Kolmogorov, who introduced the notion of descriptive complexity. In simple terms, The Kolmogorov-complexity, k(x), of an object, x, is the size of the minimum programme that encodes x. Unfortunately, Kolmogorov-complexity is non-computable: while it is conceivable to calculate it for some particular objects, it is impossible to calculate it in finite time for a generic object by computational means. As such, in general, the best that can be attained are estimates of the Kolmogorov-complexity of an object.

In psychology, the Structural Information Theory (Leeuwenberg, 1968, 1969) evolved in parallel and independently with AIT. Although SIT has been applied to different domains it is, in essence a theory about human visual perception. It can be seen as a formalization of Occam's Razor principle: the best hypothesis for a set of data is the one that leads to the largest compression, i.e. the simplest. In the context of visual perception this implies that, when interpreting visual stimuli, which by default is ambiguous, the brain prefers the simplest interpretation. A recent multidisciplinary overview of perceptual organization can be found in van der Helm (2014).

Although SIT and AIT share many similarities, namely the fact that both rely on the notion of descriptive complexity, important differences exist: (i) SIT distinguishes between structural and metrical information; (ii) the outcome of SIT is a hierarchical organization while k(x) outputs a complexity value; (iii) SIT focuses on a restricted set of regularities while AIT considers all possible regularities; (iv) due to the previous point, SIT is computable, while $k(s)$ is non-computable.

While SIT is computable, the application of SIT implies a preliminary step, encoding the artefact being measured (e.g. an image) as a symbol string where each symbol refers to a perceptual primitive (van der Helm, 2004). Thus, a symbolic representation of the artefact is required, and as such applying SIT to an image implies finding a symbolic representation to the image. As such, although insights from SIT are valuable, using SIT in the context of this work is impossible since we are dealing directly with visual stimuli (i.e. images represented in pixel format) and that the conversion of images into an adequate symbolic representation is still an open problem in computer vision. For these reasons, we use as primary source of inspiration the concepts of complexity derived from AIT, which are, in essence, common to SIT.

The size of the minimum programme that encodes an object depends not only on the object but also on the "machine" for which the programme was built. That is, by definition, complexity depends on the one encoding, or perceiving, the object. As such, and considering the purposes of the current study, one should seek Kolmogorov-complexity estimates that correlate well with complexity as perceived by humans.

The most popular image compression schemes are lossy, the encoding of the images involves a loss of detail that is, hopefully, negligible and undetectable by the human eye.

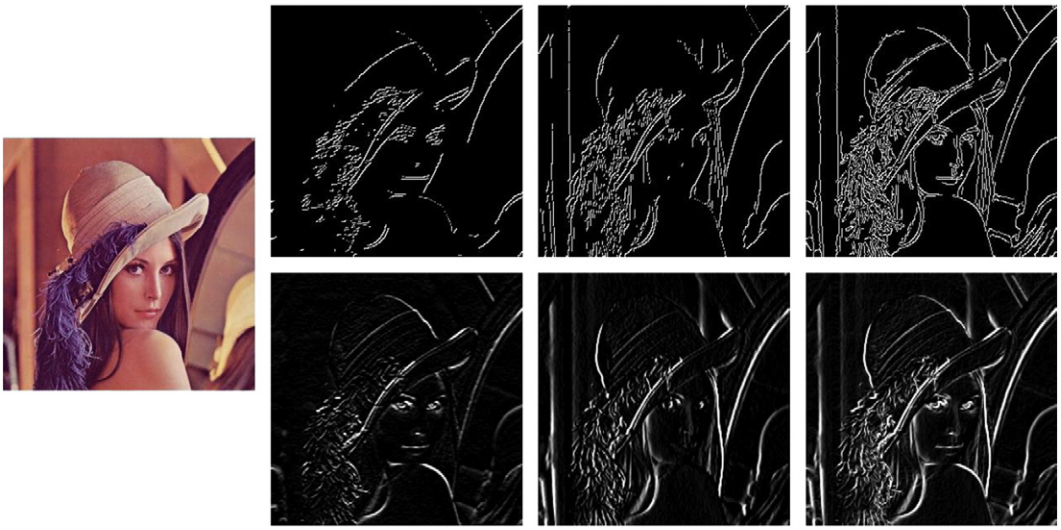

Fig. 1. Examples of the application of Canny (top row) and Sobel (bottom row) edge detection filters to a popular computer graphics testing image, "Lenna", which is presented on the left. The images illustrate, from left to right, the horizontal edges, vertical edges and all edges detected by each of the algorithms.

In this study we will use two lossy compression techniques, JPEG and fractal encoding, to estimate the perceived complexity of the images. JPEG encoding relies on the fact that human vision is much more sensitive to small variations over large areas than to the exact strength of high-frequency variations (see, e.g., Palmer, 1999). The details of the JPEG algorithm are not relevant for the scope of this paper, for the current purpose it is enough to note that: i) The JPEG encoding scheme is perceptually motivated; ii) in essence, since the image is split in blocks of 8×8 pixels, it is a low-level and local compression scheme, it does not take advantage of non-local regularities of the image. While (i) prompts its use in this study, (ii) indicates one of the most pressing shortcomings of this encoding in this context, which has driven us to consider other encoding schemes. Unlike JPEG, fractal image compression (Fisher, 1995) is a global scheme that takes advantage of the self-similarities present in the image (see Fig. 2). An image is encoded through a partitioned iterated function system, which, in simple terms, represents an image by identifying the similarities between different regions of the image and the transformations (e.g. rotation, brightness adjustment) required to produce a region from a similar one. Therefore, it takes advantage of high-level structural information and non-local regularities, which allows it to use perceptually relevant image characteristics such as symmetry, repetition, or even rhythm, to attain better compression. Due to its properties, we consider that fractal image compression is aligned with some of the principles and ideas defended by SIT. It is important to notice that fractal image compression is a general-purpose compression technique that can be applied to fractal and non-fractal images. The experimental results attained in previous studies (Machado & Cardoso, 1998; Machado & Cardoso, 2002; Machado, Romero, & Manaris, 2007; Machado, Romero, Santos, et al., 2007) indicate the adequacy of fractal image compression for estimation of image complexity.

Fig. 2. The squares indicate similarities among different areas, though many others exist.

3.3. Entropy estimations by Zipf's Law

Zipf's law (Zipf, 1949) concerns the frequency of occurrence of natural phenomena. Informally, a phenomenon follows Zipf's distribution when the frequency of occurrence is inversely proportional to its rank. Using natural language as an example, the most frequent word is, approximately, twice as frequent as the second most frequent word, three times as frequent as the third most frequent word, and so on. Zipf's law based metrics have been used with success in the musical domain (Manaris, Vaughan, Wagner, Romero, & Davis, 2003) but the results of their application in image analysis are, to the best of our knowledge, inconclusive. The usefulness of Zipf's law based metrics depends on the identification of phenomena that are perceptually relevant.

4. Objectives of the present research

Here we employ Canny and Sobel edge filters, as well as a series of estimates used in other scientific studies, to estimate the complexity of visual stimuli by computational means. In the first set of experiments (Study 1) we examine the correlation between visual complexity as rated by human participants and metrics based on JPEG compression, fractal compression and Zipf's law. The selection of these metrics was inspired by several studies which associate aesthetic with complexity (Arnheim, 1966; Machado & Cardoso, 1998; Machado, Romero, & Manaris, 2007; Machado, Romero, Santos, et al., 2007) and Zipf's law (Manaris, Purewal, & McCormick, 2002).

In the second set of experiments (Study 2) an artificial neural network is employed to predict the complexity rating assigned by humans to different images. The Artificial Neural Networks based their predictions on combinations of the metrics and filters, whose individual

performance had been correlated with human ratings in the course of the first experimental stage.

5. Study 1: correlation between perceived complexity and automatic metrics

5.1. Methods

5.1.1. Stimuli

Stimuli were taken from a set of over 1500 images digitalized and used by Cela-Conde et al. (2004, 2009). The set included abstract and representational images, some of which were artworks, while others were not. Representational and abstract stimuli differed with regard to the presence or absence of explicit content, respectively. Artistic stimuli included reproductions of renowned artists' paintings, which have been catalogued and exhibited in museums. We took paintings from different styles or movements, namely realism, cubism, impressionism, and post-impressionism. This choice was guided by the collection Movements in Modern Art of the Tate Gallery London (Cottington, 1998; Gooding, 2001; Malpas, 1997; Thomson, 1998), and supplemented with 17th and 18th European paintings. Non-artistic stimuli consisted of photographs taken from the book series Boring Postcards (Parr, 1999, 2000), a sample of images from the series of CDs Master Clips Premium Image Collection (IMSI, San Rafael, CA), used in industrial design, to illustrate books, and so on, together with some photographs taken by us. This category, thus, included artefacts, landscapes, urban scenes, and other familiar visual stimuli that would generally not be considered for exhibition in museums. Our artistic and non-artistic categories are analogous to Winston and Cupchik's (1992) distinction of high art versus popular art. They noted that whereas popular art emphasizes subject matter, especially its pleasing aspects, High art relies on a broader range of emotions and knowledge, striving to achieve a balance between content and style.

These images were either discarded or modified in order to minimize the influence of strange variables on the complexity ratings by human participants. First, aiming to avoid the impact of familiarity, only relatively unknown artworks were selected. Second, to avoid the influence of ecological variables, we eliminated those stimuli that contained clear views of human figures and human faces, or portrayed emotional scenes. Third, to reduce the undesired influence of psychophysical variables, the resolution of all stimuli was set to 150 ppi, and their size to 9 by 12 cm. Additionally, the colour spectrum was adjusted in all images. For each one, values of extreme illumination and shadow were adjusted to attain a global tone range allowing the best detail. Stimuli with a mean distribution of pixels concentrated in both the left (dark) and right (light) extremes of the histogram were discarded. Thereafter, the luminance of the stimuli was adjusted to between 370 and 390 lx. Finally, the signature was removed from all signed pictures. Any stimulus that could not be reasonably modified according to all of these specifications was discarded.

The final standardized set included 800 images in 5 categories: 262 abstract artistic (AA), 141 abstract non-artistic (AN), 149 representational artistic (RA), 48 representational non-artistic (RN), and 200 photographs of natural and human-made scenes (NHS). Examples of each category are presented in Fig. 3. The set was then divided into 8 subsets pseudo-randomly to balance stimuli categories across subsets. A group of 240 participants (112 men and 128 women, whose average age was 22.03, with a standard deviation of 3.75) was randomly divided into 8 subgroups of 30 people. Each subgroup was asked to rate the visual complexity of one of the stimuli subsets on a 1 to 5 Likert scale, ranging from very simple to very complex. The average score awarded by each corresponding subgroup of participants was considered to represent the value of perceived complexity for each stimulus in the final set. Stimuli from this set have previously been used by Cela-Conde et al. (2009), Forsythe et al. (2011), and Nadal et al. (2010).

5.2. Procedure: feature extraction

The feature extraction process implies three steps: (i) pre-processing, including all the transformation and normalization operations applied to every input image; (ii) applying filters to each image; and (iii) calculating statistical measurements and image complexity estimates.

The first step involved pre-processing the set of 800 stimuli described in the previous section. All images were individually subjected to a series of transformations before they were analysed. Each image was loaded and resized to a standard width and height of 256×256 pixels, transformed into a three channel image in the RGB (red, green and blue) colour space, with a depth of 8-bit per channel, and all pixel values scaled to the [0; 255] interval. This step ensured that all input images shared the same format and dimensions. Afterwards, each image was converted into the HSV (Hue, Saturation and Value) colour space, and its HSV channels were split. Each of these channels was stored as a 1-channel grey-scale image. From here on we will refer to these images as H, S and V. A new grey-scale image was also created by performing a pixel by pixel multiplication of S and V channels and scaling the result to [0; 255]. From here on we will refer to this image as CS (colourfulness).

The second step was applying filter operations to the images resulting from the prior one. We used two edge detection algorithms Canny (1986) and Sobel (1990), which are among the most popular edge detection algorithms in computer graphics. Both required three transformation operations: identifying horizontal edges, vertical edges, and edges in all directions. The edge detection filters were applied individually to each of the image channels mentioned previously (H, S, V, CS).

The third step was the estimation of complexity metrics. The metrics described here were individually applied to the unfiltered channel-images and to the six images resulting from the application of Canny and Sobel edge detection. The set of metrics employed can be divided into two distinct groups: generic statistical information metrics, which included average and standard deviation; complexity estimates, which included JPEG and fractal compression, Zipf rank frequency and Zipf size frequency (Machado, 2007; Machado & Cardoso, 1998; Machado, Romero, Cardoso, & Santos, 2005; Machado, Romero, & Manaris, 2007; Machado, Romero, Santos, et al., 2007).

5.2.1. Average and standard deviation

The average (Avg) and the standard deviation (StD) were calculated using the pixel intensity value of each image, except for the H-channel image. Since the H-channel is circular, the average and the standard deviation were calculated based on the norm and angle of Hue values. In addition, the Hue angle value was multiplied by the CS value, and consequentially a norm was calculated using Hue and CS values. It is important to notice that, like all other metrics, Avg and StD were applied to the filtered and unfiltered images.

5.2.2. JPEG and fractal compression

The rationale for using these methods is the following: JPEG and fractal compression are lossy compression schemes, i.e., the compressed image does not exactly match the original, producing a compression error. All other factors being equal, complex images will tend toward higher compression errors and simple images will tend toward lower compression errors. Additionally, complex images will tend to generate larger files than simple ones. Thus, the compression error and file size are positively correlated with image complexity. Considering these factors the complexity estimate of image, i, according to the lossy encoding scheme, f, is given by the following formula (Machado & Cardoso, 1998):

$$\mathrm{Complexity}(i) = \mathrm{rmse}(i,\ f(i)) \times s(f(i))/s(i),$$

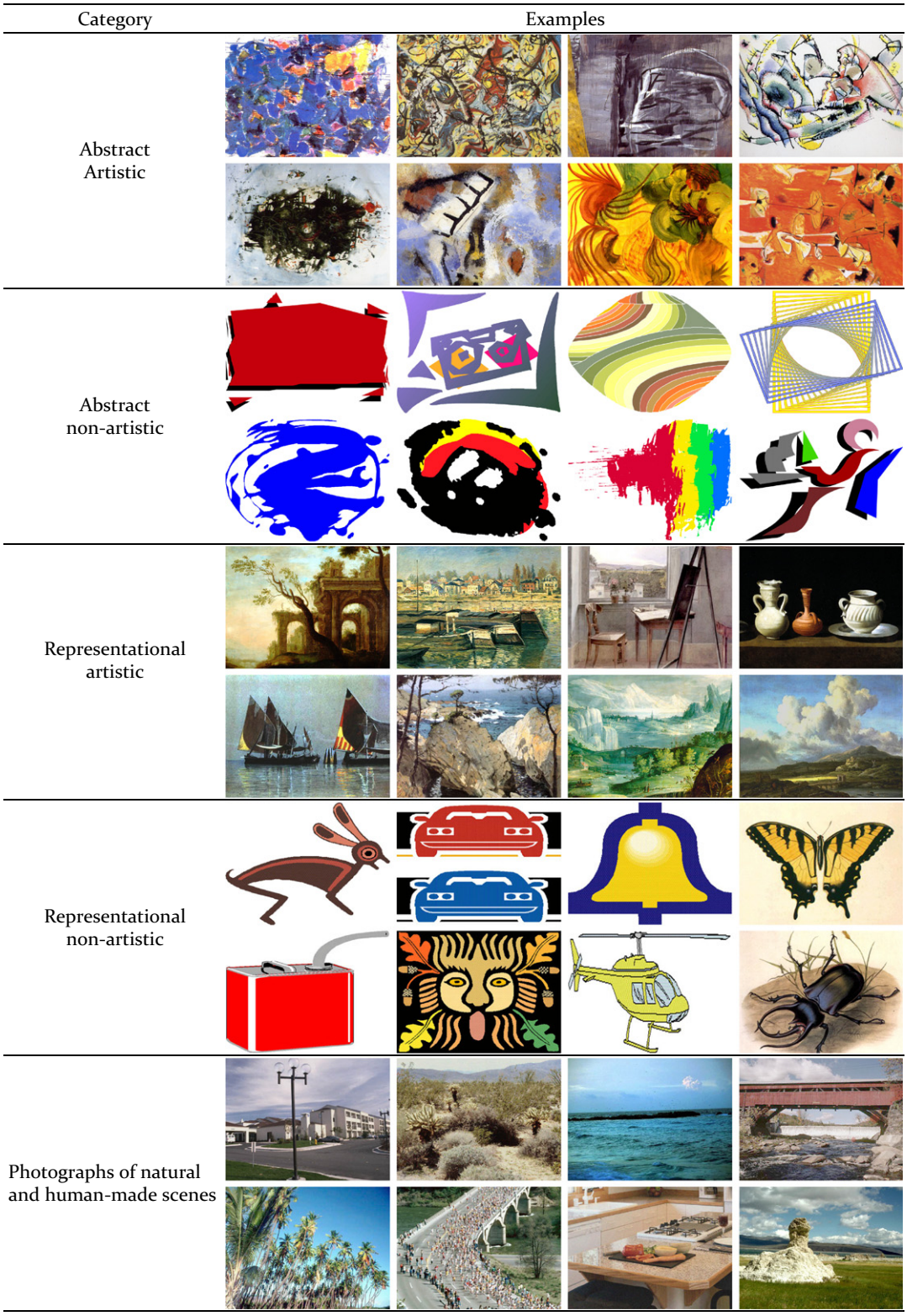

Fig. 3. Examples of stimuli of each category.

where rmse stands for root mean square error and s is the file size function.

JPEG and fractal image compression schemes allow the specification of the maximum tolerated error, which allows specifying the quality of the encoding. We considered three levels of detail for each scheme, low, medium, and high for each compression scheme. In the experiments described here, fractal compression was performed using a quad-tree fractal image compression scheme (Fisher, 1995).

5.2.3. Zipf rank frequency

Following the same rationale, and informed by work in the musical domain (Manaris et al., 2003), we also employ Zipf's law-based metrics

Table 1
Descriptive statistics for the complexity ratings awarded by humans to images in each category.

Category	n	Min.	Max.	Mean	SD
Abstract artistic	262	1.36	4.94	3.75	0.60
Abstract non-artistic	141	1.06	3.91	1.79	0.51
Representational artistic	149	1.42	4.67	3.45	0.49
Representational non-artistic	48	1.30	4.39	2.74	0.81
Natural and human-made scenes	200	1.24	4.42	2.79	0.67

(Zipf, 1949). The calculation of the Zipf rank frequency metric requires counting the number of occurrences of each pixel intensity value in the image, ordering by the number of occurrences, tracing a rank vs. number of occurrences plot using a logarithmic scale in both axis, and calculating the slope of the trend-line and the linear correlation with the trend-line (Powers, 1998).

5.2.4. Zipf size frequency

This metric was calculated in a way similar to Zipf rank frequency. For each pixel we calculated the difference between its value and the value of each of its neighbour pixels. We counted the total number occurrences of differences of size 1, size 2, …, size 255. We traced a size vs. number of occurrences plot using a logarithmic scale in both axis and calculated slope and linear correlation of the trend-line (Powers, 1998).

After the application of the metrics, the results were aggregated to form the image feature vectors. The average and standard deviation for each channel image returned two values, except for the Hue channel that returned four values for the average and two values for the standard deviation. The JPEG and fractal compression metrics returned three values each, corresponding to the three considered compression levels. Although these metrics were applied to all the images resulting from the pre-processing and filtering transformations, the JPEG metric was also applied to the RGB image. As for the Zipf's law based metrics, the slope of the trend-line (m) and linear correlation (R2) of all grey-scale images were extracted. In the case of the Hue channel, these metrics returned four values each: two considering only the Hue channel and two considering the Hue and CS channels in conjunction. The combination of pre-processing operations, filters and metrics yields a total of 329 features per image. For the sake of parsimony these features are named using a functional notation as follows: metric(filter(channel-image, <arguments>), <arguments>). For instance Fractal(Canny(S, All), High) refers to the feature resulting from the application of fractal compression to the image obtained by applying Canny edge detection, in all directions, to the saturation channel of the original image. The parameter high specifies that one is using the maximum level of detail while compressing the image.

5.3. Results

This section presents the correlations between the average complexity score awarded to each stimulus by humans (Nadal et al., 2010) and the values obtained by the computational features we propose. Throughout the paper we employ Spearman's correlation measure and, as such, from here on we will simply refer to it by using the term correlation.

Table 2 summarizes the results of this experiment by presenting, for each of the six metrics, the feature that obtained the highest correlation with the average complexity score awarded by humans to each image. As can be observed, the maximum correlations with the ratings of the entire set of images (column "All") are similar for most metrics, with JPEG(Sobel(S,All), High) attaining the maximum overall correlation, 0.771. The consistency of this set of features is highlighted by considering the correlation among them, which is always above 0.93. This minimum value is obtained by calculating the correlation between JPEG(Sobel(S,All), High) and Size(Canny(S,All), M). (See Table 1.)

In addition to the correlation to the entire set of images, Table 2 also reports the correlation for each of the five image categories. The correlations for the subset of representational artistic images are lower than for other categories, whereas the correlations for the subset of representational non-artistic stimuli are the highest.

As can be observed in Table 2, the best overall results — i.e., those where the correlation between computational estimates and human ratings is higher — were obtained by applying the edge filters to the Saturation colour channel of the images. This applies for all metrics considered in this study and constitutes, perhaps, the most striking finding of this experiment.

Fig. 4 illustrates the application of the Canny filter to three of the images used in this study. They all belong to the representational non-artistic subset, and exhibit different levels of complexity. It is clear from this figure how the higher levels of complexity are associated with a larger number and dispersion of edges, which are identified by the edge detection operations.

Table 3 reports the features that provide highest correlations with the ratings awarded by humans for each combination of filter and colour channel. As can be observed, the features using the Saturation channel yield the best results, followed by the ones using the Value channel. The results obtained when considering the Hue channel are not as reliable, which can be explained by its circular nature that leads to the consideration of non-existing edges, e.g. when the pixel values transition, directly, from 255 to 0, an edge will be detected, however, this edge does not exist since the channel is circular. Using edge detection tends to produce better results than those obtained when no filter is used (column "No Filter"). The results obtained without edge detection filters are similar to those reported by Forsythe et al. (2011).

Table 4 is similar to Table 2, however, in this case we did not employ edge detection operations. With these settings, the application of features based on JPEG compression leads to similar results to those obtained by Forsythe et al. (2011) using GIF compression estimates. Specifically, the average correlation for features using JPEG compression on the V and S channels, without filtering, is 0.701. As is the case with other metrics explored in this study, the best values are obtained when using the S channel, where a maximum correlation of 0.743 is reached. Features based on the fractal compression of the V and S channels, without filtering, yield an average correlation of 0.608 and a maximum correlation of 0.722, obtained with Fractal(NoFilter(S), High).

Considering that a negative correlation is as useful for the purposes of this study as a negative one, the best correlation obtained using Zipf

Table 2
Features exhibiting the highest correlation with human ratings across the entire set of images (column "All"). The correlation with human ratings for each of the five categories is depicted in columns AA to NHS. Numbers within parentheses indicate the number of images in each category.

Metric	Feature	All	AA (262)	AN (141)	RA (149)	RN (48)	NHS (200)
JPEG	JPEG(Sobel(S,All), High)	0.771	0.606	0.481	0.393	0.691	0.528
Fractal comp.	Fractal(Canny(S,All), High)	0.764	0.577	0.520	0.376	0.747	0.536
Zipf rank	Rank(Canny(S,All), M)	0.762	0.570	0.513	0.391	0.755	0.546
Zipf size	Size(Canny(S,All), M)	0.756	0.556	0.527	0.393	0.753	0.551
Avg.	Avg(Canny(S,All))	0.762	0.570	0.513	0.391	0.755	0.546
StD	StD(Canny(S,All))	0.762	0.570	0.513	0.391	0.755	0.546

Fig. 4. Examples resulting from applying the Canny edge detection filter to 3 images of the representational non-artistic category. The original image is shown on the left, with the average rating awarded by participants in Nadal et al.'s (2010) study. The image on the right is the result of applying the Canny filter to the saturation channel. As can be observed, higher levels of detail correspond to higher amounts and dispersion of edges. We also present the value of the *JPEG*(*Sobel*(*S*,*All*), *High*) feature for each of the images, illustrating how it correlates with the human perception of complexity.

rank frequency over the V and S channels without filtering is −0.638, corresponding to *Rank*(*NoFilter*(*S*), *R2*), while the average of the absolute values of the correlations is 0.549. Zipf size frequency yields a best correlation of 0.357 and an average correlation of 0.1363, in the same conditions.

The average and standard deviation metrics produce poor results when applied without edge detection filters to the V and S channels, the best correlation values are −0.464 and −0.286 while the average values are 0.2794 and 0.2436, respectively.

Overall these results indicate that JPEG and fractal compression metrics are robust, yielding good correlations with the average ratings assigned by humans in a wide set of conditions, even when no edge detection operation is applied. The same cannot be stated for the remaining metrics, whose performance depends, to a large extent, on the use of edge detection operations.

Table 4 also shows the average correlation with each of the five image categories. Previously, see Table 2, considering the images of the RN category provides the strongest correlations while considering

Table 3
Features exhibiting the highest correlation with the ratings awarded by humans for each combination of filter and colour channel. $Sobel_{All}$, $Sobel_{Vertical}$, $Sobel_{horizontal}$ refer to the application of the Sobel filter for detecting all edges, vertical edges and horizontal edges, respectively. The same applies for the Canny filter.

Filter/colour	$Sobel_{All}$	$Sobel_{Vertical}$	$Sobel_{Horizontal}$	$Canny_{All}$	$Canny_{Vertical}$	$Canny_{Horizontal}$	No filter
H	0.682	0.671	0.661	0.624	0.575	0.566	0.596
	$ZipfSize_M$	$Fractal_{High}$	$Fractal_{High}$	$JPEG_{Low}$	$Fractal_{Low}$	$Fractal_{Low}$	$JPEG_{High}$
S	0.771	0.764	0.761	0.766	0.737	0.719	0.743
	$JPEG_{High}$	$JPEG_{High}$	$JPEG_{High}$	$JPEG_{Low}$	$JPEG_{High}$	$JPEG_{High}$	$JPEG_{High}$
V	0.715	0.718	0.705	0.708	0.711	0.674	0.704
	$JPEG_{High}$	$Fractal_{Medium}$	$JPEG_{High}$	$JPEG_{High}$	$Fractal_{Low}$	$JPEG_{High}$	$JPEG_{High}$

Table 4
Features that do not use edge detection filters exhibiting the highest correlation with human ratings across the entire set of images. The correlation with human ratings for each of the five categories is depicted in columns AA to NHS. Numbers within parentheses indicate the number of images in each category.

Metric	Feature	All	AA (262)	AN (141)	RA (149)	RN (48)	NHS (200)
JPEG	JPEG(NoFilter(S),High)	0.743	0.569	0.374	0.377	0.638	0.500
Fractal comp.	Fractal(NoFilter(S), High)	0.722	0.554	0.467	0.377	0.622	0.499
Zipf rank	Rank(NoFilter(S), R2)	−0.638	−0.437	−0.310	−0.065	−0.596	−0.277
Zipf size	Size(NoFilter(H + CS), R2)	0.357	0.501	−0.027	−0.040	0.282	0.112
Avg.	AVG(NoFilter(V))	−0.464	−0.477	−0.164	−0.089	−0.687	−0.011
StD	STD(NoFilter(S))	−0.286	−0.099	−0.307	−0.029	−0.492	−0.057

those of the RA category provides the weakest. Therefore, although considering edge information tends to result in higher correlations, the trend is similar with or without edge information and some image categories, e.g. RA, appears to be more problematic than others in what concerns the estimation of image complexity.

Table 5 shows the correlations among the complexity estimates without using edge detection operations. As it can be observed, in general, the correlations are inferior to the ones obtained when using edge detection. The high correlation between the metrics base on image compression indicates that, although the compression methods are radically different, they tend to yield highly correlated values for most images.

5.4. Discussion

As previously mentioned the use of edge detection operators significantly improves the correlation between the computational estimates of complexity and the human ratings. This difference is particularly visible when naive metrics such as Avg. or StD are used, and becomes less accentuated when image compression metrics are employed. In general image compression techniques tend to attain higher compressions on smooth images where the value of a given pixel is predictable, in the sense that it can be estimated by considering the values of the surrounding pixels. As such, JPEG and fractal image compression will naturally tend to compress images with a low percentage of edges better than images with a high number. Since the percentage of edges influences the performance of compression, even when no edge detection operation is applied, the advantage of using edge detection is bound to be less visible for these metrics than for the others.

Estimates such as $Avg(Canny(s))$, $r_s = 0.762$, directly measure the percentage of the pixels of the image that correspond to edges. Although, this estimate shares similarities with the perimeter detection method employed by Forsythe et al. (2011) ($r_s = .58$), the experimental results indicate that, in the considered conditions, the number of edges is a better estimate of image complexity. Typically the best results are obtained using Canny edge detection, which is expected since it tends to be more reliable than Sobel edge detection. However, the highest overall correlation, $r_s = 0.771$, was obtained using Sobel edge detection and JPEG compression. Although the difference is marginal — the highest overall correlation attained with Canny edge detection coupled with JPEG compression is 0.766 — an explanation is in order. Canny edge detection produces abrupt transitions between black and white, while Sobel edge detection tends to produce smoother transitions (see Fig. 1). By design, JPEG compression deals better with smooth than abrupt transitions. As the fidelity rating increases, the difficulty in dealing with these high frequencies tends to distort the complexity estimate. Thus, although Canny is a more reliable edge detector, the images resulting from Sobel edge detection are more appropriate for JPEG compression.

The estimates based on saturation (S channel) provide the highest correlations outperforming those based on value (V channel), which is an unexpected result. In Fig. 5 we depict several examples from the dataset, presenting the original image and the results of edge detection on the S and V channels. As it can be observed, although the results tend to be similar, in some extreme cases, e.g. the first row of Fig. 5, the differences can be considerable. Overall, in the considered experimental conditions, the analysis of changes in saturation, and consequent edge detection, appears to provide a better indication of the boundaries between objects than the analysis of changes in value. Further tests are required to determine the generality of this result.

The correlations obtained using the JPEG ($r_s = .743$) and Fractal ($r_s = .722$) compression without a previous edge detection step are similar to those obtained by Forsythe et al. (2011) using GIF compression ($r_s = .74$), but higher than those reported by Marin and Leder (2013), even for their subset of IAPS images using TIFF ($r_s = .53$), JPEG ($r_s = .52$), PNG ($r_s = .46$), or GIF ($r_s = .29$) compressions.

It is interesting to notice that $avg(nofilter(v))$ yields a $r_s = -0.46$, a correlation that is higher (in absolute value) than it would be expected, since there is no apparent reason to expect that the average value would correlate negatively or positively with complexity. An analysis of the results allows us to explain this occurrence. The overall correlation of $r_s = -0.46$ results from a $r_s = -0.687$ for the stimuli of the RN category, a $r_s = -0.477$ for the AA category, and correlations close to zero for the remaining categories (see Table 4). An analysis of the images belonging to this category reveals that they are frequently composed of one or several objects positioned against a background, which is often of a light colour. Therefore, images filled with objects tend to be darker, and thus $avg(nofilter(v))$ will tend to be low, while those where the background occupies most of the space tend to be lighter, resulting in high $avg(nofilter(v))$, which explains the observed correlation.

In general, the proposed estimates perform better for representational non-artistic images, r_s up to .755 (RN); obtain intermediate results for abstract artistic (AA), abstract non-artistic (AN) and photographic (NHS) images, r_s up to 0.606, 0.527 and 0.551, respectively; and perform the worst for representational artistic images (RA), r_s up to 0.393. Reliably detecting edges of RN images tends to be a straightforward task. Although some difficult stimuli exist, the same also applies to the images in the AN and AA categories. Conversely, reliably detecting edges of RA and NHS images is, in comparison, a difficult task.

Table 5
Correlations among types of metrics. The feature showing the greatest correlation has been selected for each metric, as described in Table 4.

Metric	Feature	Human	JPEG	Fractal C.	Zipf rank	Zipf size	Avg.
JPEG	JPEG(NoFilter(S),High)	0.743					
Fractal comp.	Fractal(NoFilter(S), High)	0.722	0.984				
Zipf rank	Rank(NoFilter(S), R2)	−0.638	−0.594	−0.571			
Zipf size	Size(NoFilter(H + CS), R2)	0.357	0.301	0.287	−0.290		
Avg.	AVG(NoFilter(V))	−0.464	−0.501	−0.491	0.459	−0.485	
StD	STD(NoFilter(S))	−0.286	−0.113	−0.117	0.280	−0.375	0.384

Contrasting the results of Tables 2 and 4 confirms that the gains in performance when performing edge detection are low for the images of the RA and NHS categories.

Therefore, if we consider the reliability of edge extraction as an indicator of the quality of our metrics, then we would expect to obtain the worse results in the RA and NHS categories. As such, what remains to be explained is the difference in performance between the RA and NHS categories. Our explanation for this fact is twofold: (i) Reliably identifying edges on paintings is arguably harder than on photographs since brushstrokes may introduce low level artefacts that difficult edge detection, (ii) the semantics of the image arguably have a higher influence in the perception of complexity in representational artistic images than in photographs. Further experimentation is necessary to confirm this interpretation.

6. Study 2: prediction of image complexity using Artificial Neural Networks

6.1. Methods

6.1.1. Stimuli

Stimuli used in Experiment 2 were the same as in Experiment 1. Namely, they were 800 images divided into 5 categories: 262 abstract artistic, 141 abstract non-artistic, 149 representational artistic, 48 representational non-artistic, and 200 photographs of natural and man-made scenes (see Fig. 3 for examples of each category). Each stimulus has been rated by 30 participants, though not necessarily the same ones (see methods of Study 1, above, for further details).

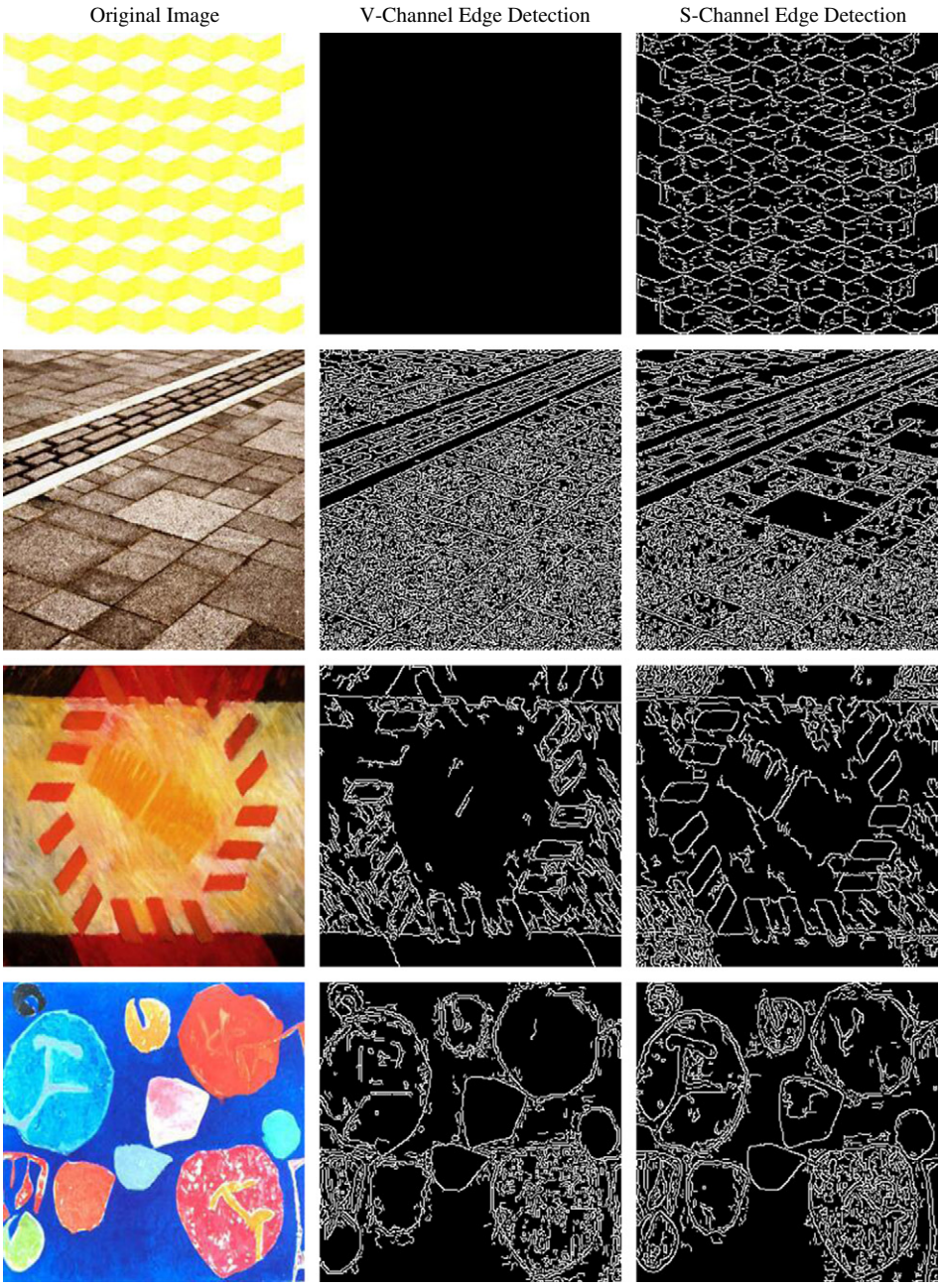

Fig. 5. Examples of stimuli belonging to the dataset (first column); the results of applying Canny edge detection to the V channel of those stimuli (second column); the results of applying Canny edge detection to the S channel of those stimuli (third column).

6.1.2. Procedure

Whereas the aim of Experiment 1 was to analyse the correlation of each individual estimate with regard to complexity as rated by humans, Experiment 2 aimed to ascertain whether the average visual complexity scores attributed by humans to each stimulus could be approximated by employing information of several estimates and a machine learning approach. As such, these experiments aim to answer two important questions: (i) Is it possible to improve the results reported in Study 1 by combining several estimates? (ii) Is it possible to use computational methods to learn to predict visual complexity as perceived by humans?

To answer these questions we conducted a series of experiments where Artificial Neural Networks (ANN henceforth) are used to predict the average visual complexity scores attributed by humans, based on the estimates described in the previous section. A thorough introduction to ANNs is beyond the scope of this paper. Haykin (1994) provides a comprehensive foundation to those interested in ANN research.

In simple terms, ANNs are computational models inspired by the brain, and are able to learn. An ANN is a system of interconnected artificial neurons, which are simplified models of biological neurons. Being one of the most popular machine learning approaches, ANNs have been used for a wide variety of tasks including: pattern recognition, clustering, classification and prediction. To perform such tasks the ANNs must be *trained* by exposing them to a set of *examples* (stimuli). Training can be *unsupervised* or *supervised*. In unsupervised training the ANNs is exposed to a set of examples and, using an appropriate algorithm, will adapt in order to minimize a cost function that should be a function of the stimuli and of ANNs' output. In supervised learning, the model used in the course of the experiments reported here, the ANN is trained by exposing it to a set of patterns each composed of a stimulus and of desired response for that stimulus. Using adaptive machine learning techniques the ANN constructs an internal model of the training patterns learning to produce the desired response for each stimulus. Successful learning requires that the ANN is able to *generalize*, i.e. producing adequate responses to patterns used for training can be accomplished through *memorization*, learning implies that the ANN is also able to produce adequate responses to patterns that were not used on its training.

The ANNs used in these experiments have a classical architecture: they are composed of an input layer of neurons, a hidden layer, and an output layer. As the names indicate, the input layer determines the data that is available to the ANN, while the output is the response of the ANN. The hidden layer is the main responsible for performing the computations required to convert the input in the desired output. Based on prior studies (Machado, Romero, & Manaris, 2007; Machado, Romero, Santos, et al., 2007; Machado et al., 2005), we empirically chose to include 15 neurons in this layer. Each image is described by means of the features presented in the Procedure section of Study 1, and the values of these features feed the input layer. Thus, the ANN does not "see" the images, it only has access to the corresponding computational visual complexity estimates. Any information that is not captured by these features is lost. Each neuron of the input layer corresponds to a feature, and vice-versa. Aiming to test different combinations of features, the input layer has a variable number of neurons. Fig. 6 shows a schematic example of the topology of ANNs that were used.

To train the ANNs we used a backpropagation algorithm (Haykin, 1994). This algorithm is based on the idea of back propagation of *error*. When exposed to a training pattern the output of the ANN is calculated. Then, this output is compared with the desired response, the difference between these values is the *error*. For the experiments described in this paper the output of the ANN is compared with the average visual complexity rating attributed by humans to the image, since this is what we are trying to approximate. The weights of the connections between neurons of the hidden and output layer are adjusted to decrease the error. In the next step, the weights of the connections between the neurons of the input and hidden layers are also adjusted to decrease the error associated to each of the neurons of the hidden layer. During training the ANN is repeatedly exposed to all training patterns until a proper response is achieved. This process implies that the ANN will progressively adjust in order to reduce the mean square error (SME henceforth) over the training set. However, we report the average error and the correlation between the response of the ANN and the visual complexity scores attributed by humans, it is important to notice, that the ANN is trained to minimize the SME, not the correlation. A better prediction from the point of view of SME can, in some circumstances, give raise to lower Spearman correlation. Therefore, improvements in terms of correlation are attained indirectly. The training parameters are presented in Table 6, and were also determined

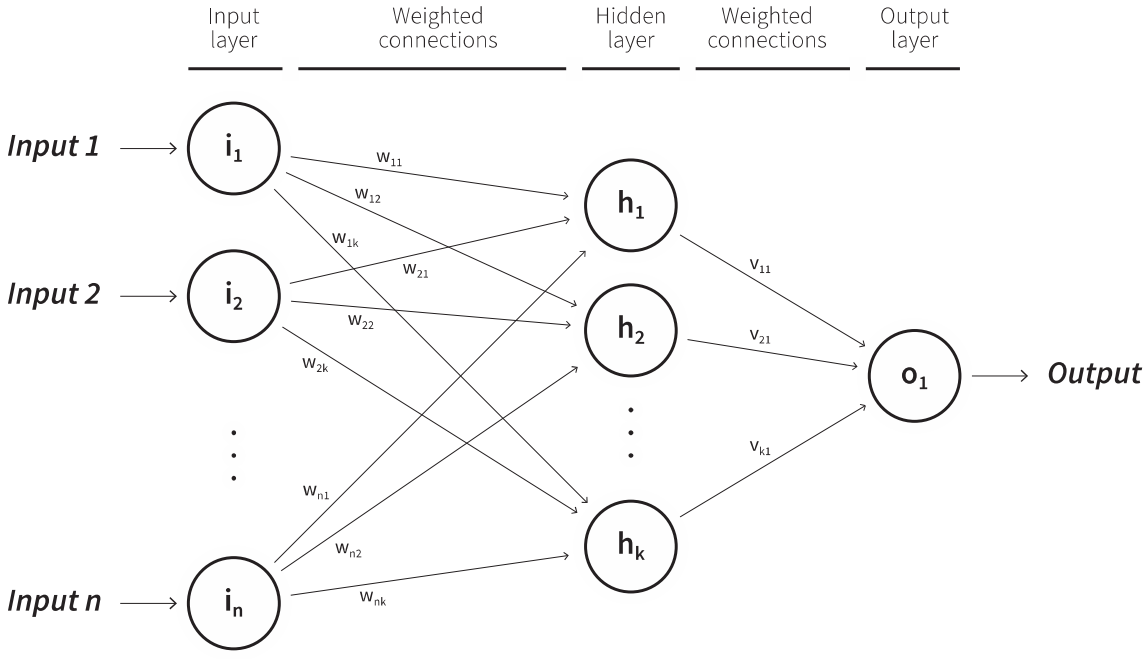

Fig. 6. Schematic representation of the ANNs' topology.

Table 6
Training parameters.

Parameters	Configuration
Cycles	250
Output function	Identity
Initial weights	Random, [−0.1; 0.1]
Training function	Back propagation
Learning rate	0.1
Admissible error	0.01
Maximum error	0.015

empirically (Machado, Romero, & Manaris, 2007; Machado, Romero, Santos, et al., 2007; Machado et al., 2005).

We used a cross-validation procedure for training and validation. This procedure involves dividing the set of 800 patterns (one for each image) into 10 sets of the same size (80 images). Training and validation of the network is performed 10 times. In each case, one of those 10 sets is used as a validation set and the other 9 as training sets. That is to say, in each repetition the network is trained using 720 patterns learning to predict the average complexity rating attributed by humans from the supplied features for these 720 patterns. After training, the adequacy of this model is validated by assessing its performance on the remaining 80 instances (which were not used for training). The process is repeated 10 times. Thus, all patterns are used once for validation and 9 times for training.

6.2. Results

In this section we describe the results obtained using different combinations of features as input of the ANNs. Based on these inputs the ANNs are trained to predict the average complexity scores awarded by the human participants. We report the average error — i.e., the average difference between the target and predicted complexity values — and the correlation between the predictions of the ANNs and the average human response. All the results reported in this paper refer to the performance obtained in validation, i.e. the performance of the ANNs on patterns that were not used in their training. The performance over training instances is, obviously, significantly higher, however, as previously explained, performance over training instances is not indicative of learning.

Table 7 reports the results obtained using 4 different combinations of features. As can be observed, the first three configurations (NET1, NET2 and NET3) achieve an average error close to 0.1 in a normalized 0 to 1 interval, which corresponds to average error of 0.4 on the 1 to 5 scale used in the study, and a correlation above 0.8. These results are better than those achievable when using individual features, indicating that the ANNs were able to (i) extract relevant information from the supplied features; (ii) combine this information in meaningful ways. As a base for comparison, and as a control experiment, we trained an ANNs using only basic metrics (mean and standard deviation) and without edge detection filters. The results obtained when using this configuration, NET4, are clearly worse than those produced by the rest of the predictors and close to correlations obtained when using the individual metrics of mean and standard deviation. These results confirm the ones presented in the previous section: as expected, these metrics are insufficient to predict visual complexity as perceived by humans.

The first of combinations, NET1, contains all the metrics, filters and colour channels extracted by the system described in the Procedure section of Study 1, amounting to a total of 329 input feature values for each image. This network leads to the best results, with an average error of 0.095 and a correlation of 0.833.

The second combination (NET2) does not take advantage of the edge detection filters. The results are slightly worse than the previous combination. They clearly demonstrate that the complexity metrics are able to extract meaningful information even when no edge detection operation is performed. The third combination (NET3), uses the edge detection filters but it does not take advantage of the complexity estimates proposed in this study (i.e. it only uses the mean and standard deviation metrics). The performance when using this configuration is comparable to the one obtained when using the NET2 configuration. The fact that NET1 yields better results than NET2 and NET3 also indicates that although there is a significant overlap between edge detection and the complexity metrics, in the sense that they capture similar types of information, this overlap is not total and that their combination provides added value.

The experiments reported in the previous section indicated that the Saturation channel was more informative than Value channel, which, in turn, was more informative than the Hue channel. Table 7 also presents the results obtained when considering the saturation channel alone. As can be observed, the results are inferior to those attainable when using the information gathered from the three colour channels. This indicates that although the Saturation channel is the most useful for predicting visual complexity, the ANNs are able to explore the information gathered from the other channels to improve their predictions.

Table 8 summarizes the performance of different ANN configurations for each of the four categories of stimuli. In terms of average prediction error the ANNs performed better on the Abstract Non-Artistic category, closely followed by the Abstract Artistic and Representational Artistic categories. The worse performances were obtained in the Representational Non-artistic and Photographs of Natural and Human-made Scenes categories. The highest correlations were observed in the Abstract Artistic and Representational Non-artistic categories, followed by the Abstract Non-Artistic, Photographs of Natural and Man-made Scenes, and Representational Artistic.

Comparing these correlations with those obtained using the *JPEG*(*Sobel*(*S*)) feature reveals that, for all categories, the NET1–NET3 configurations were able to improve the correlation with human's perception of visual complexity through the exploration of the information provided by different features. This improvement is particularly visible for the Abstract Artistic and Abstract Non-Artistic categories — which also obtained the lowest average prediction errors.

It is interesting to notice that although the correlation results for Representational Non-artistic stimuli is the highest, the average error is also high. This result, although unexpected, can be easily explained: Representational Non-artistic stimuli only account for 6% of the total number of stimuli (48 out of 800), therefore the ANNs have not been

Table 7
Results obtained with different combinations of features.

Colour channels	With complexity metrics		Without complexity metrics	
	With filters	Without filters	With filters	Without filters
All colour channels	*NET1*	*NET2*	*NET3*	*NET4*
	Input: 329	Input: 47	Input: 70	Input: 10
	Spearman: 0.833	Spearman: 0.806	Spearman: 0.809	Spearman: 0.471
	Avg. error: 0.095	Avg. error: 0.103	Avg. error: 0.103	Avg. error: 0.152
Saturation	*NET5*	*NET6*	*NET7*	*NET8*
	Input: 84	Input: 12	Input: 14	Input: 2
	Spearman: 0.791	Spearman: 0.782	Spearman: 0.766	Spearman: 0.343
	Avg. error: 0.104	Avg. error: 0.108	Avg. error: 0.110	Avg. error: 0.169

Table 8
Performance of the different ANN configurations for each category of stimuli in terms of average error and correlation. The correlations obtained using the *JPEG*(*Sobel*(*s*)) feature are shown on the right for comparison purposes.

	Average error				Correlation				
	NET1	NET2	NET3	NET4	NET1	NET2	NET3	NET4	JPEG(Sobel(S))
All (800)	0.095	0.103	0.103	0.152	0.833	0.806	0.809	0.471	0.771
AA (262)	0.076	0.081	0.077	0.107	0.750	0.739	0.729	0.477	0.606
AN (141)	0.074	0.075	0.078	0.088	0.597	0.585	0.481	0.328	0.481
RA (149)	0.087	0.087	0.086	0.092	0.398	0.397	0.432	0.163	0.393
RN (48)	0.099	0.088	0.099	0.166	0.753	0.779	0.717	0.372	0.691
NHS (200)	0.104	0.112	0.110	0.138	0.587	0.573	0.581	0.016	0.528

exposed to sufficient examples to fine-tune their predictions for this category. This interpretation is confirmed by the fact that the NET2 configuration, which uses fewer inputs, obtained better results than the NET1 configuration. Thus, in simple terms, the number of training patterns is insufficient for the ANNs to learn how to take advantage of the additional data and hence performance decreases (it is well-established that, in general, as one increases the number of inputs the number of training patterns should also increase to allow the ANNs to explore the additional information).

Although the average error for Representational Artistic images is low, the correlation results for this category are the worst, and there is not a significant improvement in correlation over the results obtained with the feature that yield the highest correlation. Thus, although the ANNs were able to predict the visual complexity of these stimuli with reasonably high precision, they are unable to accurately rank them. Our interpretation is that the considered set of features is not sufficient to accurately rank these stimuli in terms of complexity and we hypothesise that this is linked with the influence that the semantics associated with these stimuli may have on human's perception of visual complexity. Our explanation for the results obtained for the Photographs of Natural and Human-made Scenes category are, in essence, similar.

6.3. Discussion

The configuration that yields the best overall results both in terms of average prediction error and correlation with the average visual complexity scores attributed by humans is NET1. This is an expected result since this configuration uses all available information — i.e. the 329 features — as input. The NET2 and NET3 configurations obtain comparable results but use different data as input. NET2 uses all computational complexity estimates but no edge detection filters while NET3 uses edge detection filters and the average and standard deviation metrics. Although this indicates that there is a significant overlap among the information provided by the inputs for NET2 and NET3 — i.e. that the different features are assessing the same information by different means — the fact that NET1 outperforms both indicates that the overlap is not total.

The configurations that only use information gathered from the saturation channel, NET5 to NET8, obtain worse performance that those that have access to the features gathered from the three colour channels. Showing that, although the saturation channel is the most informative for the prediction of visual complexity, the additional information present in the Value and Hue channels can be used to improve the ANNs' predictions. The results highlight the intrinsic limitations of processing the Hue channel, which is circular, with the set of metrics and filters proposed in this study.

Overall, the results of Study 2 demonstrate that it is possible to improve upon the correlation results obtained when using individual features by combining the information provided by different features. Moreover, they demonstrate that it is possible to use Machine Learning techniques, ANNs in this case, to learn to combine the information provided by different features, which is an important result.

The computational prediction of the average visual complexity scores attributed by humans to each stimulus by means of ANNs is one of the novel contributions of this study. The average prediction error using the best ANN configuration was 0.095 in a normalized 0 to 1 interval, which corresponds to an average prediction error of 0.4 in the 1 to 5 scale used when gathering human responses. Given the variability of human's responses we consider this prediction error to be acceptable.

These results become particularly relevant when the experimental conditions are taken into consideration. Typically, to attain good results, ANNs with a number of inputs of this magnitude should be trained with thousands (or tens of thousands) of training patterns. Thus, training sets of 720 patterns are far from ideal. Additionally, the training patterns correspond to four different categories of stimuli and the number of stimuli belonging to each category varies significantly. Ideally the number of stimuli of each class should be equal. Having unbalanced training sets can significantly hinder learning due to the underexposure of the ANNs stimuli of the classes with lower cardinality. This effect is visible in the analysis of the results described in the previous section concerning the performance of the ANNS on different categories.

Considering these limitations, it becomes reasonable to infer that improvements, both in terms of average prediction error and correlation, could be attained if one was using a training set without this shortcomings and specifically designed for Machine Learning purposes.

7. General discussion

Over the last decade, the interest in developing robust computational measures of visual complexity has gained momentum (e.g. Donderi, 2003, 2006; Forsythe et al., 2003, 2008, 2011; Marin & Leder, 2013; Palumbo, Ogden, Makin, & Bertamini, 2014). This line of research is driven by two interrelated goals. The first of these is to produce accurate predictions of humans' impression of visual complexity of objects, scenes, or designs. For basic psychological research, such measures have the potential to significantly reduce time and costly resources (e.g. participants, laboratory space and materials) invested in preparing sets of visual stimuli to obtain complexity pre-ratings, to select images representing a given range of complexity (e.g. Cela-Conde et al., 2004, 2009), and so forth. For applied research, such measures could provide straightforward, fast, accessible, and easily implemented indications of people's aesthetic and affective responses to products, devices, and designs, as well as their behaviour, and ability to use and interact with them (Bauerly & Liu, 2008; Krishen et al., 2008; Lavie et al., 2011; Reimann et al., 2010).

The second goal of this line of research is to understand the psychological construction of perceived complexity, in terms of the visual features and cognitive processes it relies on. Computational measures contribute to this research with two complementary questions: (i) why do human and computational measures agree to the extent that they do? (ii) Why do they disagree to the extent that they do? In answer to the first question, the computational measures of complexity can be used to model the human response to complexity, under the assumption that highly accurate predictors indicate processes (e.g. edge detection) or features (e.g. contrast, saturation) that are relevant to both human- and computer-generated complexity values. In answer to the second question, the computational measures can be used to

identify the factors that bias humans' experience of visual complexity away from the complexity present in the image. There is evidence suggesting that the same image can be experienced as more complex or simpler depending on its familiarity (Forsythe et al., 2008), artistic appearance (Forsythe et al., 2011; Marin & Leder, 2013; Nadal et al., 2010), content (Marin & Leder, 2013), or context (Tinio & Leder, 2009). That is to say, the experience of visual complexity is constructed on the grounds of visual features present in the object, but not solely. The perceiver's knowledge, experience, and understanding of the visual object also come into play. Why this happens—even when participants are requested to strictly focus on the visual features of complexity—and how this happens, are questions that remain open to research.

Obviously, the success of the aforementioned goals rests on the quality of the computational estimations of complexity. In the two studies presented here, we have examined the performance of a series of metrics and edge detection operations in estimations of human participants' visual complexity of images. The metrics were based on image compression error and Zipf's law. For edge detection we relied on Canny (1986) and Sobel (1990) filters. The experiments involved 800 images belonging to 5 different categories. Thirty participants had previously rated each of these images on perceived complexity.

In a first set of experiments we assessed the correlation between individual computational features and human's perception of complexity. The experimental results indicate that edge detection, even when coupled with naive metrics, yields strong correlations with human's perception of visual complexity (see Results section of Study 1). Taking into account the relevance of edge detection in the early stages of (human) visual perception, we consider that this result provides insights for the neurological basis of the perception of visual complexity.

When no edge detection operations were used, the only metrics that provided satisfying correlations were the ones based in image compression. The obtained results ($r_s = .743$) are comparable to those obtained by Forsythe et al. (2011) when using GIF compression ($rs = .74$), further attesting the viability of using image compression techniques in this context. Our results share, however, somewhat different from those reported by Marin and Leder (2013). In the case of their IAPS subset of stimuli, most of the correlations range between $r_s = .46$ and $r_s = .53$, but in the case of their set of artworks, there were no significant correlations between complexity ratings made by human participants and any of the image compression metrics. The stimuli used in the experiments reported in the present study are the same as those used by Forsythe et al. (2011), and different to those used by Marin and Leder (2013). This suggests that, with regard to highly complex artistic depictions, the correlation between image compression and human ratings might be sensitive to the process used to assemble the stimuli set, including the way complexity is defined and manipulated (Nadal et al., 2010), the range of complexity represented in the set, the inclusion or exclusion of images depicting affective scenes varying in arousal and pleasantness, or the inclusion or exclusion of images depicting human themes. Further studies are required to determine whether such procedural options modulate the performance of the metrics, the human ratings, or both.

The estimates based on saturation (S channel) provided the highest correlations, closely followed by those based on value (V channel). This result is unexpected and further testing is required to determine if it is generalizable or if it results from the specificities of the experimental settings.

In a second set of experiments we used machine learning techniques to learn to predict visual complexity from a set of computational features. The experimental results show that, through machine learning, it is possible to combine the information provided by several features to produce better estimates than those obtained when using individual features ($r_s = 0.833$ vs. $r_s = 0.771$). Moreover, the average difference between the predictions of the system and the average visual complexity scores attributed by humans to each stimulus is below 0.4 in a 1 to 5 interval (0.095 in a normalized 0 to 1 interval). As far as we are aware of, this is the first study where machine learning techniques are used to predict human's perception of visual complexity.

We tested our machine learning system with different combinations of inputs, concluding that the one that performed better was the one that had access to all of the proposed features. The results obtained when using edge detection and naive features are comparable to those obtained when image compression without edge detection. This indicates that there is a significant overlap in the information gathered by these features, but also that the overlap is not total since using all the features yields better results.

An analysis of the performance of the machine learning system on different categories of stimuli reveals that the difficulties in accurately predicting the visual complexity of Representational Artistic images and Photographs of Natural and Man-made Scenes may be linked with the influence that the semantics of the image has on human's perception of visual complexity.

As previously mentioned (see Discussion section of Study 2) the use of machine learning techniques recommends the availability of a significantly larger number of stimuli, which should also be distributed equally among the different categories. Further testing will address these limitations by considering well-balanced and broader set of stimuli. Likewise, features based on image salience, texture analysis and colour distribution, among others, can also provide additional information and contribute to better estimates.

To conclude, in this study we have presented ways to estimate people's perception of visual complexity in scenes with greater accuracy than previous ones (e.g. Forsythe et al., 2011). Our results indicate that edge density and compression error are the best predictors of participants' complexity ratings, suggesting that the perceptual and cognitive processes involved in detecting edges and dealing with non-redundant information play crucial roles in the subjective experience of complexity. Nevertheless, this experience seems to be influenced, to a certain extent, by semantic aspects, leading to variations in the accuracy of predictions depending on image category (e.g. photograph, artwork). Finally, as shown by our machine learning study, the most accurate predictions are produced via the combination of multiple image features, suggesting that the perception of visual complexity emerges from the interaction of several dimensions (Berlyne et al., 1968; Nadal et al., 2010).

References

Alario, F. -X., & Ferrand, L. (1999). A set of 400 pictures standardized from French: Norms for name agreement, image agreement, familiarity, visual complexity, image variability, and age of acquisition. *Behavior Research Methods, Instruments, & Computers*, *31*, 531–552.

Arnheim, R. (1966). *Towards a psychology of art/entropy and art—An essay on disorder and order*. The Regents of the University of California.

Attneave, F. (1957). Physical determinants of the judged complexity of shapes. *Journal of Experimental Psychology*, *53*, 221–227.

Bauerly, M., & Liu, Y. (2008). Effects of symmetry and number of compositional elements on interface and design aesthetics. *International Journal of Human Computer Interaction*, *24*, 275–287.

Berlyne, D.E. (1963). Complexity and incongruity variables as determinants of exploratory choice and evaluative ratings. *Canadian Journal of Psychology*, *17*, 274–290.

Berlyne, D.E. (1970). Novelty, complexity, and hedonic value. *Perception & Psychophysics*, *8*, 279–286.

Berlyne, D.E. (1971). *Aesthetics and psychobiology*. New York: Appleton-Century-Crofts.

Berlyne, D.E. (1974). Novelty, complexity, and interestingness. In D.E. Berlyne (Ed.), *Studies in the new experimental aesthetics: Steps toward an objective psychology of aesthetic appreciation* (pp. 175–180). Washington, D. C.: Hemisphere Publishing Corporation.

Berlyne, D.E., Ogilvie, J.C., & Parham, L.C.C. (1968). The dimensionality of visual complexity, interestingness, and pleasingness. *Canadian Journal of Psychology*, *22*, 376–387.

Bertamini, M., Palumbo, L., Gheorghes, T.N., & Galatsidas, M. (2015). Do observers like curvature or do they dislike angularity? *British Journal of Psychology*. http://dx.doi.org/10.1111/bjop.12132 (in press).

Birkhoff, G.D. (1932). *Aesthetic measure*. Cambridge, Mass.: Harvard University Press.

Bonin, P., Peereman, R., Malardier, N., Méot, A., & Chalard, M. (2003). 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, & Computers*, *35*, 158–167.

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-8*(6).

Cela-Conde, C.J., Ayala, F.J., Munar, E., Maestú, F., Nadal, M., & Capó, M.A. (2009). Sex-related similarities and differences in the neural correlates of beauty. *Proceedings of the National Academy of Sciences of the United States of America, 106*, 3847–3852.

Cela-Conde, C.J., Marty, G., Maestú, F., Ortiz, T., Munar, E., & Fernández, A. (2004). Activation of the prefrontal cortex in the human visual aesthetic perception. *Proceedings of the National Academy of Sciences of the United States of America, 101*, 6321–6325.

Chatterjee, A. (2004). Prospects for a cognitive neuroscience of visual aesthetics. *Bulletin of psychology and the arts, 4*, 55–60.

Cottington, D. (1998). *Cubism.* London: Tate Gallery Publishing.

Cupchik, G.C. (1986). A decade after Berlyne. New directions in experimental aesthetics. *Poetics, 15*, 345–369.

Donderi, D.C. (2003). *A complexity measure for electronic displays: Final report on the experiments.* Toronto: Department of National Defence, Defence Research and Development Canada.

Donderi, D.C. (2006). Visual complexity: A review. *Psychological Bulletin, 132*, 73–97.

Donderi, D.C., & McFadden, S. (2005). Compressed file length predicts search time and errors on visual displays. *Displays, 26*, 71–78.

Eysenck, H.J. (1941). The empirical determination of an aesthetic formula. *Psychological Review, 48*, 83–92.

Eysenck, H.J. (1942). The experimental study of the 'Good Gestalt' — A new approach. *Psychological Review, 49*, 344–363.

Eysenck, H.J., & Castle, M. (Aug 1971). Comparative study of artists and nonartists on the Maitland Graves design judgment test. *Journal of Applied Psychology*, *55*(4), 389–392. http://dx.doi.org/10.1037/h0031469.

Fechner, G.T. (1876). *Vorschule der Ästhetik.* Leipzig: Breitkopf und Härtel.

Fisher, Y. (Ed.). (1995). *Fractal image compression: Theory and application*. London: Springer Verlag.

Forsythe, A., Mulhern, G., & Sawey, M. (2008). Confounds in pictorial sets: The role of complexity and familiarity in basic-level picture processing. *Behavior Research Methods, 40*, 116–129.

Forsythe, A., Nadal, M., Sheehy, N., Cela-Conde, C.J., & Sawey, M. (2011). Predicting beauty: Fractal dimension and visual complexity in art. *British Journal of Psychology, 102*, 49–70.

Forsythe, A., Sheehy, N., & Sawey, M. (2003). Measuring icon complexity: An automated analysis. *Behavior Research Methods, Instruments, & Computers, 35*, 334–342.

García, M., Badre, A.N., & Stasko, J.T. (1994). Development and validation of icons varying in their abstractness. *Interacting with computers, 6*, 191–211.

Gooding, M. (2001). *Abstract art.* London: Tate Gallery Publishing.

Graves, M. (1948). *Design judgement test.* New York: The Psychological Corporation.

Haykin, S. (1994). *Neural networks: A comprehensive foundation.* Prentice Hall PTR.

Heath, T., Smith, S.G., & Lim, B. (2000). Tall buildings and the urban skyline. The effect of visual complexity on preferences. *Environment and Behavior, 32*, 541–556.

Imamoglu, C. (2000). Complexity, liking and familiarity: Architecture and non-architecture Turkish students' assessments of traditional and modern house facades. *Journal of Environmental Psychology, 20*, 5–16.

Jones-Smith, K., & Mathur, H. (2006). Fractal analysis: Revisiting Pollock's drip paintings. *Nature, 444*, E9–E10.

Krishen, A., Kamra, K., & Mac, F. (2008). Perceived versus actual complexity for websites: Their relationship to consumer satisfaction. *Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior, 21*, 104–123.

Lang, P.J., Bradley, M.M., & Cuthbert, B.N. (2005). International affective picture system (IAPS): Affective ratings of pictures and instruction manual. *Technical report* (pp. A −7). Gainesville, FL: University of Florida.

Lavie, T., Oron-Gilad, T., & Meyer, J. (2011). Aesthetics and usability of in-vehicle navigation displays. *International Journal of Human-Computer Studies, 69*, 80–89.

Lavie, T., & Tractinsky, N. (2004). Assessing dimensions of perceived visual aesthetics of web sites. *International Journal of Human-Computer Studies, 60*, 269–298.

Leder, H., Belke, B., Oeberst, A., & Augustin, D. (2004). A model of aesthetic appreciation and aesthetic judgments. *British Journal of Psychology, 95*, 489–508.

Leeuwenberg, E.L.J. (1968). *Structural information of visual patterns: An efficient coding system in perception.* The Hague: Mouton.

Leeuwenberg, E.L.J. (1969). Quantitative specification of information in sequential patterns. *Psychological Review, 76*, 216–220.

Machado, P. (2007). *Inteligncia Artificial e Arte.* (Ph.D. thesis) Coimbra, Portugal: University of Coimbra (in Portuguese).

Machado, P., & Cardoso, A. (1998). Computing aesthetics. In F. Oliveira (Ed.), *XIVth Brazilian Symposium on Artificial Intelligence SBIA'98. LNAI Series*. (pp. 219–229). Porto Alegre, Brazil: Springer.

Machado, P., & Cardoso, A. (2002). All the truth about NEvAr. *Applied Intelligence, Special Issue on Creative Systems, 16*(2), 101–119.

Machado, P., Romero, J., Cardoso, A., & Santos, A. (2005). Partially interactive evolutionary artists. *New Generation Computing, 23*(42), 143–155.

Machado, P., Romero, J., & Manaris, B. (2007). Experiments in computational aesthetics: An iterative approach to stylistic change in evolutionary art. In J. Romero, & P. Machado (Eds.), Springer Berlin Heidelberg.

Machado, P., Romero, J., Santos, A., Cardoso, A., & Pazos, A. (2007). On the development of evolutionary artificial artists. *Computers & Graphics, 31*(6), 818–826.

Malpas, J. (1997). *Realism.* London: Tate Gallery Publishing.

Manaris, B., Purewal, T., & McCormick, C. (2002). Progress towards recognizing and classifying beautiful music with computers—Midi-encoded music and the zipfmandelbrot law. *Proceedings of the IEEE southeastcon 2002 conference, Columbia.*

Manaris, B., Vaughan, D., Wagner, C., Romero, J., & Davis, R.B. (2003). Evolutionary music and the Zipf–Mandelbrot Law: Progress towards developing fitness functions for pleasant music. *EvoMUSART2003—1st European Workshop on Evolutionary Music and Art, Essex, UK. Lecture Notes in Computer Science, Applications of Evolutionary Computing, LNCS, 2611*. (pp. 522–534). Springer.

Marin, M., & Leder, H. (2013). Examining complexity across domains: relating subjective and objective measures of affective environmental scenes, paintings and music. *PLoS One, 8*(8), e72412. http://dx.doi.org/10.1371/journal.pone.0072412.

McDougall, S.J.P., Curry, M.B., & de Bruijn, O. (1999). Measuring symbol and icon characteristics: Norms for concreteness, complexity, meaningfulness, familiarity, and semantic distance for 239 symbols. *Behavior Research Methods, Instruments, & Computers, 31*, 487–519.

McDougall, S.J.P., de Bruijn, O., & Curry, M.B. (2000). Exploring the effects of icon characteristics on user performance: The role of icon concreteness, complexity, and distinctiveness. *Journal of Experimental Psychology: Applied, 6*, 291–306.

Moshagen, M., & Thielsch, M.T. (2010). Facets of visual aesthetics. *International Journal of Human-Computer Studies, 68*, 689–709.

Nadal, M., Munar, E., Marty, G., & Cela-Conde, C.J. (2010). Visual complexity and beauty appreciation: Explaining the divergence of results. *Empirical Studies of the Arts, 28*, 173–191.

Palmer, S.E. (1999). *Vision Science: Photons to Phenomenology.* MIT Press. 978-0-262-16183-1.

Palmer, S.E., Schloss, K.B., & Sammartino, J. (2013). Visual aesthetics and human preference. *Annual Review of Psychology, 64*, 77–107.

Palumbo, L., Ogden, R., Makin, A.D.J., & Bertamini, M. (2014). Examining visual complexity and its influence on perceived duration. *Journal of Vision, 14*, 1–18. http://dx.doi.org/10.1167/14.14.3.

Parr, M. (1999). *Boring postcards.* London: Phaidon Press.

Parr, M. (2000). *Boring postcards USA.* London: Phaidon Press.

Pecchinenda, A., Bertamini, M., Makin, A.D.J., & Ruta, N. (2014). The pleasantness of visual symmetry: Always, never or sometimes. *PLoS One, 9*, e92685. http://dx.doi.org/10.1371/journal.pone.0092685.

Pieters, R., Wedel, M., & Batra, R. (2010). The stopping power of advertising: Measures and effects of visual complexity. *Journal of Marketing, 74*, 48–60.

Powers, D. (1998). Applications and explanations of Zipf's law. *NeMLaP3/CoNLL '98: Proceedings of the joint conferences on new methods in language processing and computational natural language learning* (pp. 151–160). Morristown, NJ, USA: Association for Computational Linguistics.

Reimann, M., Zaichkowsky, J., Neuhaus, C., Bender, T., & Weber, B. (2010). Aesthetic package design: A behavioral, neural, and psychological investigation. *Journal of Consumer Psychology, 20*, 431–441.

Rosenblatt, F. (Nov 1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review, 65*(6), 386–408.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning internal representations by error propagation. In D.E. Rumelhart, J.L. McClelland, & PDP Research Group (Eds.), *Paralled Distributed Processing. Explorations in the Microstructure of Cognition. Foundations, 1*. (pp. 318–362). Cambridge, MA: The MIT Press.

Salomon, D. (1997). *Data Compression: The Complete Reference.* New York, NY, USA: Springer-Verlag New York, Inc.

Simon, H.A. (1972). Complexity and the representation of patterned sequences of symbols. *Psychological Review, 79*, 369–382.

Snodgrass, J.G. (1997). Picture naming by young children: Norms for name agreement, familiarity, and visual complexity. *Journal of Experimental Child Psychology, 65*, 171–237.

Sobel, I. (1990). *An isotropic 3 × 3 image gradient operator.* Machine vision for three-dimensional scenes, 376–379.

Spehar, B., Clifford, C.W.G., Newell, B.R., & Taylor, R.P. (2003). Universal aesthetic of fractals. *Computers & Graphics, 27*, 813–820.

Strother, L., & Kubovy, M. (2003). Perceived complexity and the grouping effect in band patterns. *Acta Psychologica, 114*, 229–244.

Tatarkiewicz, W. (1972). The great theory of beauty and its decline. *The Journal of Aesthetics and Art Criticism, 31*, 165–180.

Taylor, R.P., Micholich, A.P., & Jonas, D. (1999). Fractal analysis of Pollock's drip paintings. *Nature, 399*, 422.

Taylor, R.P., Micholich, A.P., & Jonas, D. (2002). The construction of Jackson Pollock's fractal drip paintings. *Leonardo, 35*, 203–207.

Thomson, B. (1998). *Post-impressionism.* London: Tate Gallery Publishing.

Tinio, P.P.L., & Leder, H. (2009). Just how stable are stable aesthetic features? Symmetry, complexity, and the jaws of massive familiarization. *Acta Psychologica, 130*, 241–250.

van der Helm, P.A. (2004). Transparallel processing by hyperstrings. *Proceedings of the National Academy of Sciences of the United States of America, 101*(30), 10862–10867.

van der Helm (2014). *Simplicity in vision: A multidisciplinary account of perceptual organization.* Cambridge University Press.

Winston, A.S., & Cupchik, G.C. (1992). The evaluation of high art and popular art by naive and experienced viewers. *Visual Arts Research, 18*, 1–14.

Zipf, G.K. (1949). *Human Behavior and the Principle of Least Effort.* Addison-Wesley.