Visual complexity modelling based on image features fusion of multiple kernels

Carlos Fernandez-Lozano[1], **Adrian Carballal**[1], **Juan Romero**[1], **Antonino Santos**[1], **and Penousal Machado**[2]

[1]Computer Science Department. Faculty of Computer Science, University of A Coruña, A Coruña, 15071, Spain
[3]CISUC, Department of Informatics Engineering, University of Coimbra, 3030 Coimbra, Portugal

Corresponding author:
Adrian Carballal[1]

Email address: adrian.carballal@udc.es

ABSTRACT

Determining measures of human's visual complexity is one of the most important and challenging problems with respect to cognitive and affective processes. The average complexity scores attributed by two hundred and forty participants to approximately 800 visual stimuli by employing information of several estimates and different machine learning approaches, are studied in this work. Each image is rated by thirty participants according to the perceived complexity. Every stimulus is described by 329 features based on image compression error and Zipf's Law over three color channels. The performance of some representative state-of-the-art Machine Learning approaches are described by means of the R-squared correlation value and the Root Mean Squared Error. An exhaustive outlier analysis is used to determine extreme images and how they could affect the obtained regression value. The results with our Machine Learning models are of great relevance, in accordance with the psychological findings of the human conception of visual complexity.

INTRODUCTION

The development of computational measures to determine the visual complexity is a line of research which has been pursued in recent years. Finding a robust predictor of humans' impressions in terms of visual complexity could be applied to many fields, such as design, advertisement, and art, among other visual disciplines. Determining the perceptual, cognitive or contextual features that influence visual complexity would allow to anticipate human's aesthetic response in many fields.

First approaches proposed measures based on a counting system whereby elements (lines and angles) and the regularity, irregularity, and heterogeneity of those elements additively contribute to a mathematical calculation of visual complexity Birkhoff (1933); Eysenck and Castle (1971a,b). In 1932, Birkhoff Birkhoff (1933) formulated complexity as M = O/C, where "M" is aesthetic measure or value, "O" is aesthetic order, and "C" is complexity. In other words, beauty increases as complexity decreases. Thereafter, Eysenck and Castle Eysenck and Castle (1971b) studied the correlation between Birkhoff's measure and complexity measuring fifty figures against their seven-point-scale aesthetic pleasantness judgements from 1100 participants (100 artists and 1000 non-artists). Only very slight differences were observed between the experimental and control groups according to these authors. On the other hand, Lempel and Ziv Lempel and Ziv (2006) developed an algorithm to measure visual complexity. Their algorithm was based on the smallest computer program required to store/produce an image as the basis for the compression techniques we use today. The idea is based on the theory that the minimum length of the code required to describe an image is an adequate measure of complexity Leeuwenberg (1969).

Inspired by this concept, Donderi Donderi (2006) argued that the adequacy of image compression techniques to predict subjective complexity was directly related to the Algorithmic Information Theory. According to Aksentijevic and Gibson, the "algorithmic complexity is defined in terms of the length of the shortest algorithm in any programming language, which computes a particular binary string" Aksentijevic

and Gibson (2012). Aiming to address this approach, various edge detection methods such as Perimeter Detection, Canny, and others, based on phase congruency have been shown to be a reliable way of measuring complexity in the visual domain Forsythe et al. (2011); Marin and Leder (2013).

The most popular and widely used method to determine visual complexity is to derive a set of images and ask some participants to rate their complexity Cycowicz et al. (1997); Alario and Ferrand (1999). Following this methodological line, Forshyte at al. Forsythe et al. (2011) examined the performance of a series of metrics related to JPEG 2000, GIF compression and perimeter detection over 800 visual stimuli evaluated by 240 humans who provide ratings with a bound and previously indicated range of complexity Cela-Conde et al. (2009). According to the authors, GIF compression was correlated most strongly with human judgments of complexity for 800 artistic and nonartistic, abstract and representational images. Their results showed that this computational measure was significantly correlated with judged complexity getting a R-Squared = 0.5476.

In this context, Marin and Leder Marin and Leder (2013) compared several computational measures correlated with participants' complexity ratings of different kinds of materials. They found that TIFF file size (R-Squared = 0.2809) and JPEG file size (R-Squared = 0.2704) correlated strongest with subjective complexity ratings rather than measures of perimeter detection using a subset of stimuli selected from the International Affective Picture System. Differences between the results obtained by Forsythe et al. and Marin and Leder may have to do with both materials and procedure.

Recently, Machado et al. Machado et al. (2015) have proposed a wide range of new possible complexity estimates. These features were based on image compression error and Zipf's law Zipf (1949). Every feature was calculated for each image by applying different edge detection filters over all color channels. For edge detection, authors examined the performance of the well-known filters Sobel Sobel (1990) and Canny Canny (1986). Consequently, a total of 329 features based on seven metrics applied over three color channels (Hue, Saturation and Value), and using both above-mentioned filters, were extracted Guyon et al. (2006).

According to Machado et al. Machado et al. (2015), estimates which share similarities with the perimeter detection method employed by Forsythe et al. Forsythe et al. (2011), which directly measure the percentage of the pixels of the image that correspond to edges, obtain better results than those from the state-of-the-art. Otherwise, similar metrics to those obtained by Forsythe et al. Forsythe et al. (2011) using GIF compression, based on JPEG and Fractal compression error and no edge detection application, obtain similar results. Nevertheless, features directly related to the measurement of the number of edges are a better estimate of image complexity (R-Squared = 0.5806) than the perimeter detection method employed by Forsythe et al. Forsythe et al. (2011). Finally, they stated that the highest overall correlations were obtained using JPEG compression, after previously applying Canny (R-Squared = 0.5868) and Sobel (R-Squared = 0.5944) edge detector filters.

Machado et al. Machado et al. (2015) stated that edge density and compression error were 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.

Furthermore, analyzing the correlation of individual estimate, they tried to evaluate the results by combining their feature vector and Artificial Neural Networks (ANN). Their aim was to improve the results by the use of a computational method to predict human's complexity scores. They reported an increase of R-Squared = 0.6939 using all proposed features as inputs and an MLP. Despite the fact that this result upturned the correlation regarding a unique metric, it should be noted that the difference entailed the use of 329 features instead of a single one. They tested with different combinations of inputs, concluding that the one that performed best was the one that had access to all the proposed features.

Considering such a proposal as an initial approach, the choice of a model based on ANN is not always the best choice in tasks exclusively related to prediction. Neural networks in general and Multilayer Perceptrons in particular tend to present overfitting, especially with few samples Lawrence et al. (1997). There are methods to minimize such problems, such as the use of the *dropout* technique or combining the predictors of many different models Hinton et al. (2012) but none of them was applied. Therefore, in this study, other models based on Machine Learning from the latest state-of-the-art applied to the same input data used by Machado et al. Machado et al. (2015) are proposed. The objective is not only to improve existing results, but also to conduct a statistically more rigorous study, which may confirm the presence of outliers and their relevance in the process of regression.

2/15

MATERIALS AND METHODS

The authors have tested the different computational models using 10-fold cross-validation to split the data and 50 runs per model in order to evaluate the performance across different experiments and study the standard deviation. The performance of the models is evaluated using R-squared (R^2) and Root Mean Squared Error (RMSE) as well as the number of features (metrics).

0.1 Stimulus selection

In the field of aesthetic psychology there have been numerous works in which human subjects have been used to evaluate different images following aesthetic criteria (Forsythe at al. Forsythe et al. (2011), Martinez et al. Martinez and Benavente (1998), Amirshahi et al. Amirshahi et al. (2014)). For example, Schettino et al. Schettino et al. (2016) recruited seventeen male (age in range 19--33) to evaluate one-hundred pictures depicting various everyday scenes (e.g., people at the supermarket, at the restaurant, playing music, or doing sport), as well as nude female bodies and heterosexual interactions, were selected from the IAPS Lang et al. (2008).

In Street et al. Street et al. (2016), four hundred and forty-three participants, 228 men and 204 women, aged between 17 and 88 evaluate 81 abstract monochrome fractal images (9 full sets of 9 iterations of FD) generated using the mid-point displacement technique.

Besides, Lissenko et al. Lyssenko et al. (2016), 19 participants (19–37 years old) chosed 79 images from the collection of 150 images of abstract artworks that was compiled by Hayn-Leichsenring et al. (2017). By last, Friedenberg and Liby (2016) selected twenty-five undergraduates (5 men and 20 female) from Manhattan College in New York to evaluate 10 images determined by authors.
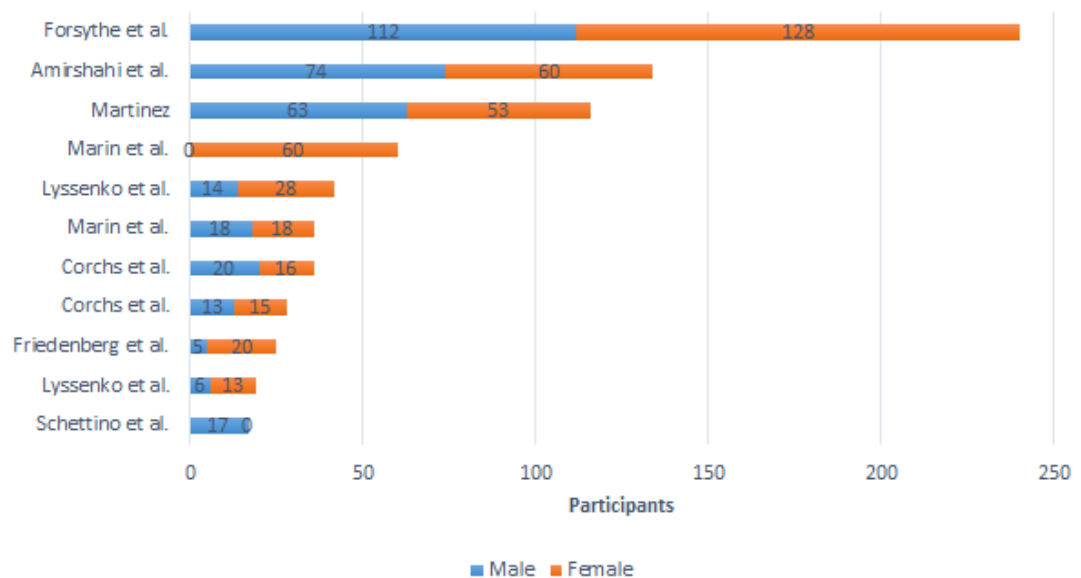
Figure 1. Number and sex of the participants in state-of-the-art analysed works

An overview of all datasets revised is shown attending, the number and sex of the participants in Table 1 and attending the available stimuli used in experimentation in Table 2. Taking both into account, Forsythe's et al. Forsythe et al. (2011) seems to be adequate for this research in terms of participants and stimuli used.

0.2 Forsythe's et al. stimuli

Stimuli were initially provided by Cela-Conde et al. Cela-Conde et al. (2009), containing a set of over 1500 digitilized images including abstract and representational images and artworks. Representational and abstract stimuli difference relies on the presence or absence of explicit content. Artistic stimuli included reproductions of renowned artists' paintings, all catalogued and exhibited in museums. The authors took paintings from different styles during 19th an 20th century: realism, cubism, impressionism,
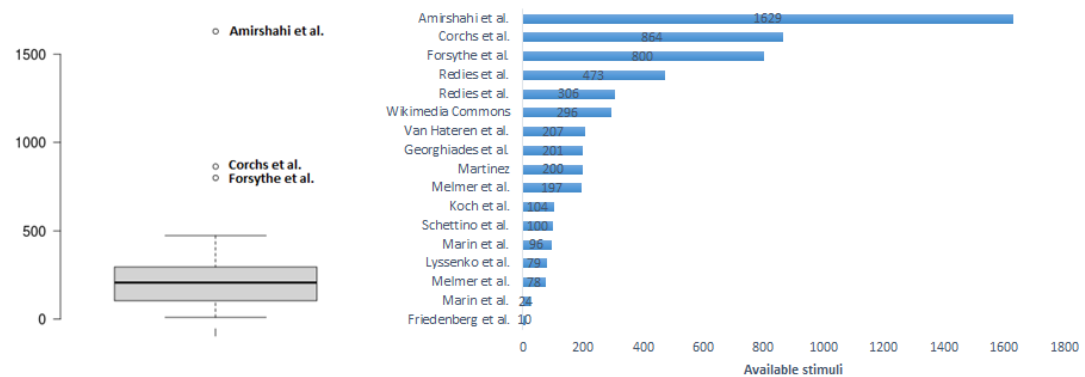
Figure 2. Avaliable images comparison (right) and datasets of stimuli (left) in state-of-the-art analysed works

and postimpressionism. Non-artistic stimuli were gathered from different book series and collections such as Boring Postcards Parr (1999, 2000) and CDs Master Clips Premium Image Collection (IMSI, San Rafael, CA). This category included artefacts, landscapes, urban scenes, and others considered of interest for its exhibition in museums. Artistic and non-artistic categories were defined analogous to Winston and Cupchik's Winston and Cupchik (1992) distinction method of popular art versus high art. According Cela-Conde et al. 'popular art emphasizes subject matter, especially its pleasing aspects, high art relies on a broader range of knowledge and emotions'.

Aiming to avoid the impact of familiarity some images were either discarded or modified. Images containing clear views of human figures or faces, or portrayed emotional scenes were also eliminated. Stimuli with a mean distribution of pixels concentrated in both extremes of the histogram were discarded. All these modifications were focus on minimize the influence of strange variables.

Additionally, all stimuli was set to 150 ppi, and their size to 9 by 12 cm and the color spectrum was adjusted in all images to reduce the influence of influence of psychophysical variables and the luminance of the stimuli was adjusted to between 370 and 390 lx. In some cases, author's signature was removed manually for proper anonymization.

The final standardized set included 800 images grouped into 5 categories: abstract artistic (AA), abstract non-artistic (AN), representational artistic (RA), representational non-artistic (RN), and photographs of natural and human-made scenes (NHS).

0.3 Participants

According to Forsythe et al. Forsythe et al. (2011), two hundred and forty participants (112 men and 128 women) from the University of the Balearic Islands, without formal artistic training, took part in the study.

The 800 images were divided into 8 equal distributed sets of 100 images. At the beginning of the experiment each participant was given a definition of complexity Snodgrass (1997) as reference. Then, participants were seated between 2 and 7 meters from a visual display where 100 images were presented with ratio 16:9 and size 400x225cm. Each image was displayed for 5 second. All participants rated these pictures on a scale from 1 to 5, being 5 catalogued as very complex and 1 as very simple.

In Figure 3 the scatter plot of the assessments made by men and women is shown. In this case, the existing R-squared correlation is 0.8485. This value could be understood as the maximum level achievable in this particular problem.

0.4 Computational models

The authors performed several experiments in order to select the best model using the R package R Core Team (2016) and MATLAB®. Some of the used computational models looked for the smallest subset of variables of the original set which provide a better performance Blum and Langley (1997), or at least equal to that obtained when using all the possible variables, considering this is a Feature Selection (FS) approach Fernandez-Lozano et al. (2015); Saeys et al. (2007); Bolón-Canedo et al. (2013); Jain and Zongker (1997). There are mainly three different approaches for FS known as filter Hall and Smith (1999), wrapper Kohavi and John (1997) and embedded.
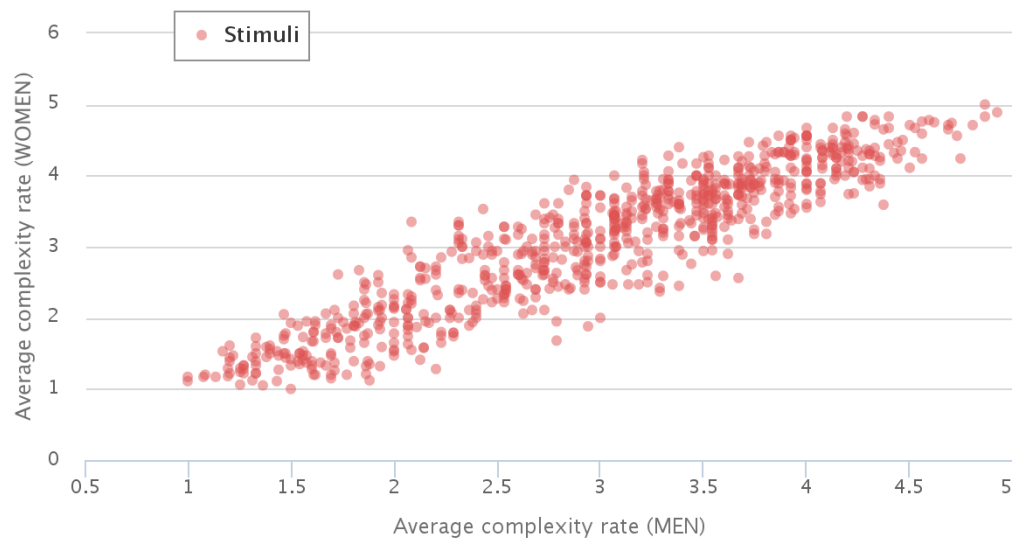
Figure 3. Complexity correlation between men and women's rating of stimuli (Source: Forsythe et al. Forsythe et al. (2011))

More specifically, the used methods are as follows: a recently proposed Feature Selection Multiple Kernel Learning (FSMKL) Fernandez-Lozano et al. (2016b) which is a filter approach for classification tasks Dash and Liu (1997), the well-known Support Vector Machines - Recursive Feature Elimination (SVM-RFE) Guyon et al. (2002); Chang and Lin (2011); Maldonado et al. (2011), Elastic Net (ENET) Zou and Hastie (2005), Lasso Tibshirani (1996) which includes embedded approaches, Generalized Linear Model with Stepwise Feature Selection (GLM) Hocking (1976) which selects features that minimizes the AIC score and the most basic standard Multiple Linear Regression (LM) without FS. The capabilities of the RRegrs Package Tsiliki et al. (2015) were enhanced in order to implement the SVM-RFE, ENET, Lasso, GLM and LM and to avoid finding the best model according to the proposed methodology as, according to García et al. (2010) it should be done based on a null hypothesis test. This package was also enhanced in order to avoid the initial splitting process, and an external cross-validation process was performed to avoid selection bias as suggested by Ambroise and McLachlan (2002), and the last step was modified in order to easily extract the results for all the models. The FSMKL was also improved following the criteria in Rakotomamonjy et al. (2008) in order to solve regression problems Menden et al. (2017). Next, the original proposed ranking criterion, proposed by Guyon Guyon et al. (2002) for SVM-RFE, was enhanced using w^2 to measure the importance of each feature.

Finally, some other R packages were used in different parts of this study, more specifically: Caret from Jed Wing et al. (2016), Kernlab Karatzoglou et al. (2004), cvTools Alfons (2012), doMC Analytics and Weston (2015), car Fox and Weisberg (2011) and ggplot2 Wickham (2009).

In order to design our experiments, a novel methodology for development of experimental designs was applied in regression problems with multiple machine learning regression algorithms Fernandez-Lozano et al. (2016a).

RESULTS

The final set of experimental stimuli was composed of 800 images grouped into 5 categories: AA, AN, RA, RN and NHS. Six different Machine Learning computational models were used in order to evaluate visual complexity. Some of them were complex approaches that evaluated the dataset following a feature selection approach.

The models previously published in Machado et al. (2015) were used as baseline for comparison purposes with our proposal. In this work, authors identified as the best single feature for this particular problem the one that calculated the JPEG compression error in the saturation color channel, applying edge detection filters previously. With respect to the ANN used in this work, they were specific for four different configurations: NET1 contained all the metrics filters and extracted color channels, NET2 did

201 not make the most of the edge detection filters, NET3 used the edge detection filters, but did not make the
202 most of the proposed complexity estimates, whereas NET4 used only basic metrics (mean and standard
203 deviation), without edge detection filters. Those five state-of-the-art results are shown in Fig. 4 with lines
204 of different colors.
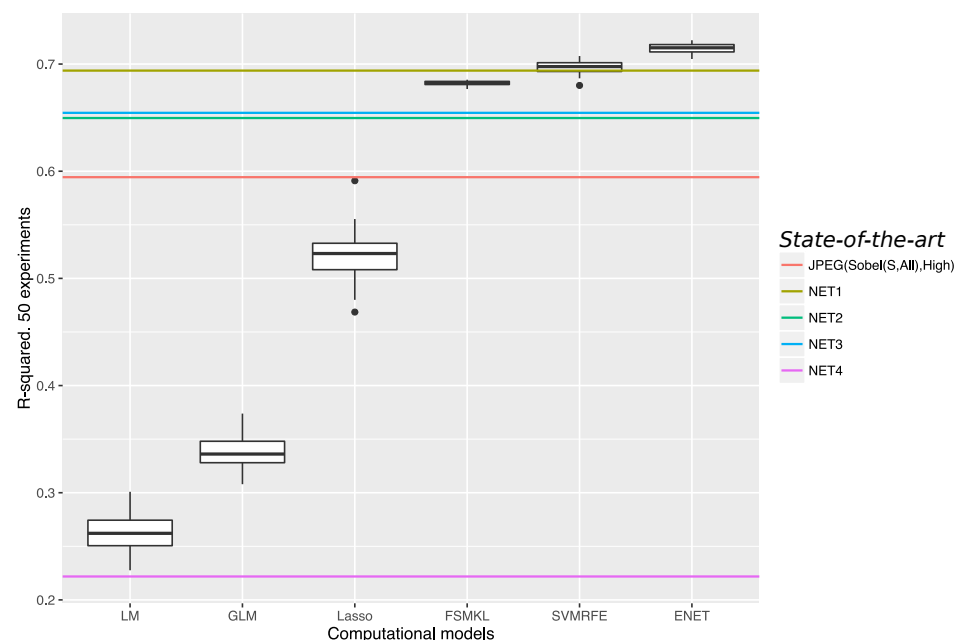
Figure 4. R-squared comparison between the state-of-the-art results and the six computational methods used in this work

205   The experiments started using six ML techniques for the visual complexity regression problem in
206 order to compare the results with those previously obtained by Machado et al. (2015). In this paper,
207 using all color channels, authors reported four Artificial Neural Networks (ANN) with R^2 values ranging
208 from 0.6938 to 0.2218. As shown in Fig. 4 SVM-RFE and ENET outperformed the best published
209 results, FSMKL achieved results very close to the previously published ones, whilst LM, GLM and Lasso
210 obtained poor results.

211   From the best three models, the most stable one in terms of R-squared, RMSE and number of features
212 was the novel FSMKL as shown in Fig. 5. An MKL approach was aimed at simultaneously learning
213 a kernel and the associated parameters. In the current study, classical kernels were used: Gaussian
214 (with sigma values of 0.1, 0.2, 0.3, 0.4, 0.5, 1 and 2) and polynomial (degrees 1, 2, 3 and 4). Thus,
215 the importance (weight) of each kernel can be measured in the final solution. Furthermore, as the use
216 of FSMKL was proposed,firstly all the features were ranked following a filter approach and thus, our
217 searching space included for each group of features (with size n), one kernel per each size value of
218 features (ranging from 2 to n). That means that the same feature could be available in different kernels
219 with different weight value, and the particular sum of weights could be calculated for each feature in the
220 final solution.

221   Finally, FSMKL used the same twenty-two features in forty-one out of the fifty experiments (see
222 Table 1). It should be noted that from these twenty-two features (see Table 1), none corresponded to
223 Machado's base metrics related to average and standard deviation of pixel values. In addition, all the best
224 features identified in Machado et al. (2015) by family complexity, except those already mentioned, were
225 used by the FSMKL.

226   According to Machado et al. (2015), the Saturation color channel was more informative than the Value
227 channel, which, in turn, was more informative than the Hue channel. According to the data shown in
228 Table 1, two aspects should be pointed out: the first is that the Hue channel was not used by any kernel in
229 FSMKL, and the second, that the weights in some cases associated with both Saturation and Value are
230 similar, such as Feature 1 and 2, and Features 6 and 7 from Table 1.

231   Similarly, all features related to JPEG and Fractal compression-based models were used in the above-
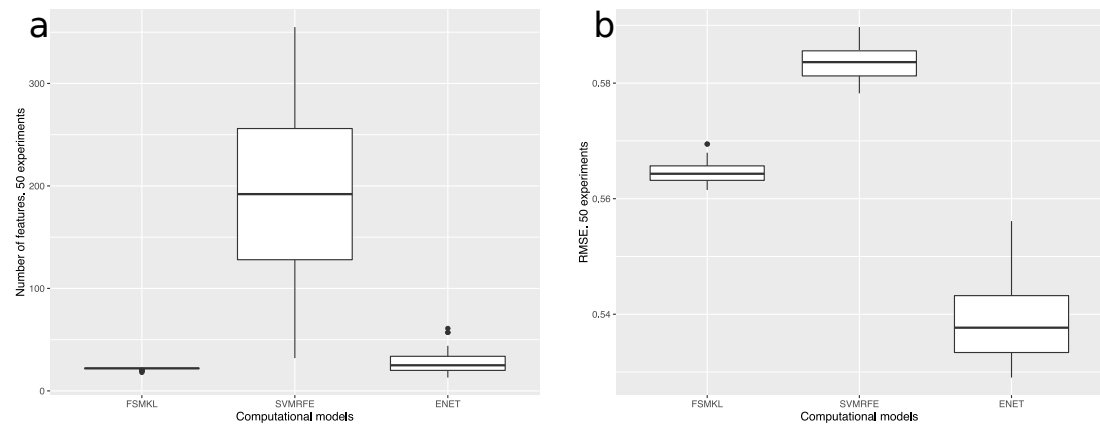
Figure 5. a) Number of features and b) RMSE of the best three computational models used in this work.

Table 1. Set of features used by FSMKL, organized by importance. The identified features were: the position (POS), the accumulated weight and the numbers of kernels in which they were used. All the features were identified using the terminology proposed in Machado et al. (2015)

Pos	Feature	Weight	Kernels	Pos	Feature	Weight	Kernels
1	Fractal(NoFilter(S),High)*	2.627	3	12	JPEG(Canny(S),High)*	0.747	4
2	Fractal(NoFilter(V),High)	2.627	3	13	JPEG(Canny(S),Medium)	0.747	4
3	Fractal(NoFilter(V),Medium)	2.216	2	14	JPEG(NoFilter(S),Medium)	0.483	2
4	Rank(NoFilter(S),R2)*	2.122	4	15	Fractal(Canny(S),High)*	0.393	2
5	Rank(NoFilter(S),M)	2.122	4	16	Size(Canny(S),M)*	0.358	4
6	JPEG(NoFilter(S),High)*	1.335	5	17	Size(Canny(V),High)	0.358	4
7	JPEG(NoFilter(V),High)	1.335	3	18	JPEG(Canny(S),Low)	0.326	4
8	Size(NoFilter(S),M)	1.195	4	19	Size(NoFilter(V),M)	0.149	2
9	Size(NoFilter(H+CS),R2)*	1.195	4	20	Size(NoFilter(V),R2)	0.149	2
10	Fractal(Canny(S),Low)	0.883	3	21	Rank(Canny(S),R2)	0.128	4
11	Fractal(Canny(S),Medium)	0.883	3	22	Rank(Canny(S),M)*	0.128	1

* Features previously identified by Machado et al.Machado et al. (2015) as the best individual features for solving the problem.

mentioned channels. Although there is a very high internal and external correlation between both families of features, FSMKL uses these minimal differences to optimize the objective function.

It is also interesting that the features calculated without using edge detection filters are those which were increasingly important when obtaining the regression model. In Machado et al. (2015) it was observed that the difference between JPEG and Fractal was not significant. In contrast, features related to Rank Frequency and Size Frequency reached high cumulative weights and were used by various kernels, such as Fractal and JPEG.

Figure 6 shows the predominance of the metrics related to the calculation of the compression error and the main use of the S channel, as observed in Machado et al. (2015). On the other hand, although the features obtained using filters have previously given better results individually, when combined, it seems that those which did not use the filters obtained higher regression coefficients.

Due to the high stability of FSMKL and to the low number of features, it was decided to analyze possible outliers using this method, as the aim of a feature selection process is to find the lowest number of features that performs equally or ideally better than the full set. Technically, FSMKL is a filter feature selection approach and SVM-RFE and ENET are embedded approaches Saeys et al. (2007).

Furthermore, after a comparison study of experimental analysis, it was concluded that FSMKL recently proposed for texture analysis in two-dimensional gel electrophoresis, along with SVM-RFE and ENET were, according to a null hypothesis test, statistically better than the others for the visual complexity problem. All the results reported refer to the performance obtained in validation using 10-fold cross-validation and 50 independent runs.
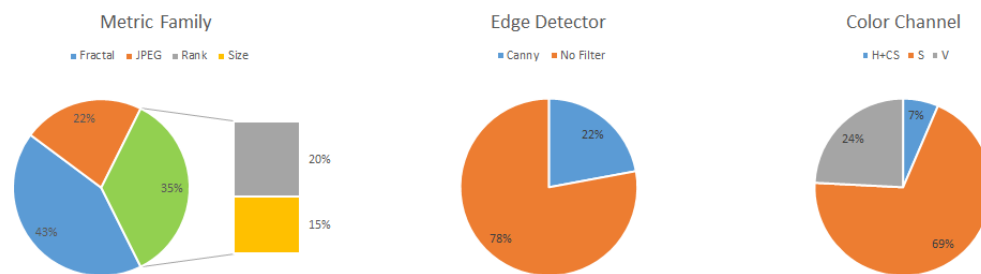
Figure 6. Prevalence of features relating to metric family (left), edge detection (center) and HSV color channel (right) of the twenty-two recurrent features according to the FSMKL method.

0.5 Outlier removal

In our opinion, some images had extreme values that could have affected the regression training of the algorithms. After training the system, an outlier analysis (Cook's distance, studentized residuals and high leverage points) was performed in order to detect the samples that could be considered extreme outliers, and thus change the fit of the model. Furthermore, a diagnosis of regression analysis was carried out using diagnostic residual scatterplots of the Pearson residuals against the fitted values. The main objective of the four graphical methods in Figure 7 is to assess the adequacy of the regression model and to find the possible outliers that generate problems and a decrease in performance R. Dennis Cook (1997). A Tukey's test was performed with the data from Figure 7.a with the null hypothesis that the model is additive, and a p-value of 0.611 was obtained, which is not lower than the significance level $\alpha = 0.05$ and the null hypothesis could not be rejected. In order to identify the influence of a particular image on the regression coefficient, the partial regression plot in Figure 7.b was plotted. Furthermore, the normality of the data in Figure 7.c was checked and an outlier test was performed using the Bonferroni correction with the null hypothesis that there were outliers in our data, thus the null hypothesis with a value of 0.5988 could not be rejected. Figure 7.d shows a bubble-plot combining the display of Studentized residuals, hat-values, and Cook's distance (represented by the size of the circles).

After these analyses, it was decided to remove six of the images as shown in Figure 8, and the algorithms were trained again, without these outliers. An improvement was observed in the R-squared cross-validation score in this new set of other fifty experimental runs. Figure 9.a shows a comparison between the best three models in Figure 4 with and without outliers. Under the removed images in Figure 8a-8e, the following was plotted:

- the influence of each particular image on the final model according to Cook's distance measure

- the Studentized residuals (residuals divided by their estimated standard deviation)

- the p-values of the outliers test (Bonferroni correction)

- the hat-values in order to identify the influential observations

Regarding the visual content of the outliers, there were several aspects to consider. It should be noted that five out of six detected outliers belonged to the NHS group. In this respect, it should be taken into account that these outliers were well distributed, as it can be observed in the AVG values from Figures 8a-8f. In Machado et al. (2015), the authors concluded that this particular group of images differed in several different ways with respect to the other four, probably because it was more difficult to detect edges in such images.

On the other hand, in the case of the first outlier (see Figure 8a) belonging to the AN group (abstract non-artistic), it may even seem reasonable to be detected as an outlier. This image, given its composition, as well as the nature of the metrics used, was interpreted as a repetitive texture or pattern, which may have been misleading to the system.

Surprisingly, the most initially promising technique, ENET, decreased its performance dramatically once the outliers were removed. At this point, it was decided to review previous works using ENET and it was found that was initially developed to overcome the limitations of Lasso, and in general outperformed

8/15
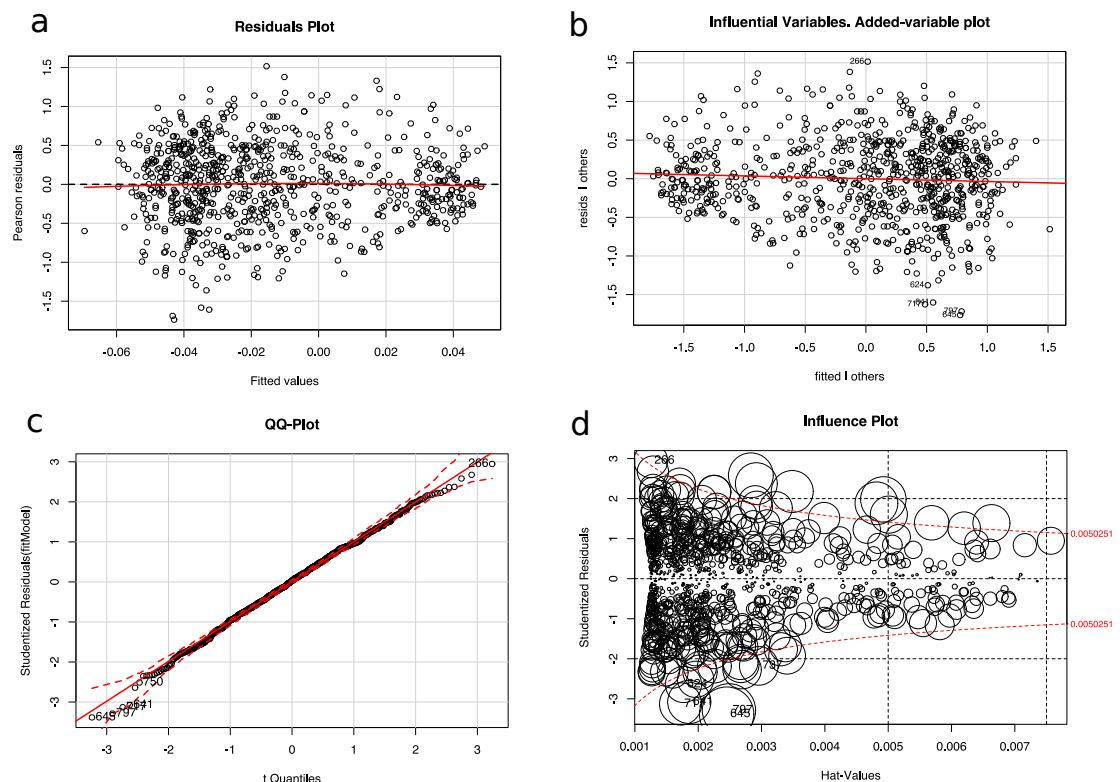
Figure 7. Diagnosis of the regression analysis: a) Residuals plot of fitted vs Pearson residuals b) Influential variables in an added-variable plot c) Normality qq-plot and d) Influence plot (Studentized residuals by hat values, with the areas of the circles representing the observations proportional to Cook's distances)

Lasso when the predictors were highly correlated Zou and Hastie (2005); Waldmann et al. (2013), as it had grouping effects and tended to select a group of highly correlated variables, once one variable was selected among them Tibshirani (1996); Bhlmann and van de Geer (2011). Furthermore, ENET tended to overfit the data in general, but more precisely when it was able only to build clusters of covariables of small size. This occured because of the ENET attempts to explain the variation adding small noise variables to the clusters Do et al. (2013), and this may seem plausible in our dataset, as there were 48 groups of correlated features with 7.8 ± 2.5 features each. Finally, the correlated variables were redundant in the sense that no additional information was obtained by adding them and sometimes they may insert noise in the clusters Guyon et al. (2002).

However, SVM-RFE reduced the whiskers although it seemed that a new outlier appeared in its boxplot (see Figure 9) and FSMKL was able to dramatically increase its performance without the outliers, reducing at the same time the variance between the experiments, but at the expense of the fact that some of them were no longer quartiles, as they became outliers. The other three methods also improved their results, or obtained similar ones, thus, it may seem that, given the lack of a set of unknown images to re-validate all models, it seems clear that the overfitting trend showed by ENET occurred with the images in our dataset.

(a) Outlier 1

Class=AN
AVG = 2.36
Index = 266

(b) Outlier 2

Class=NHS
AVG = 2.48
index = 624

(c) Outlier 3

Class=NHS
AVG = 3.55
index = 641

(d) Outlier 4

Class=NHS
AVG = 1.61
index = 645

(e) Outliner 5

Class=NHS
AVG = 3.18
index = 717

(f) Outlier 6

Class=NHS
AVG = 2.48
index = 797

(g) Outlier analysis using FSMKL
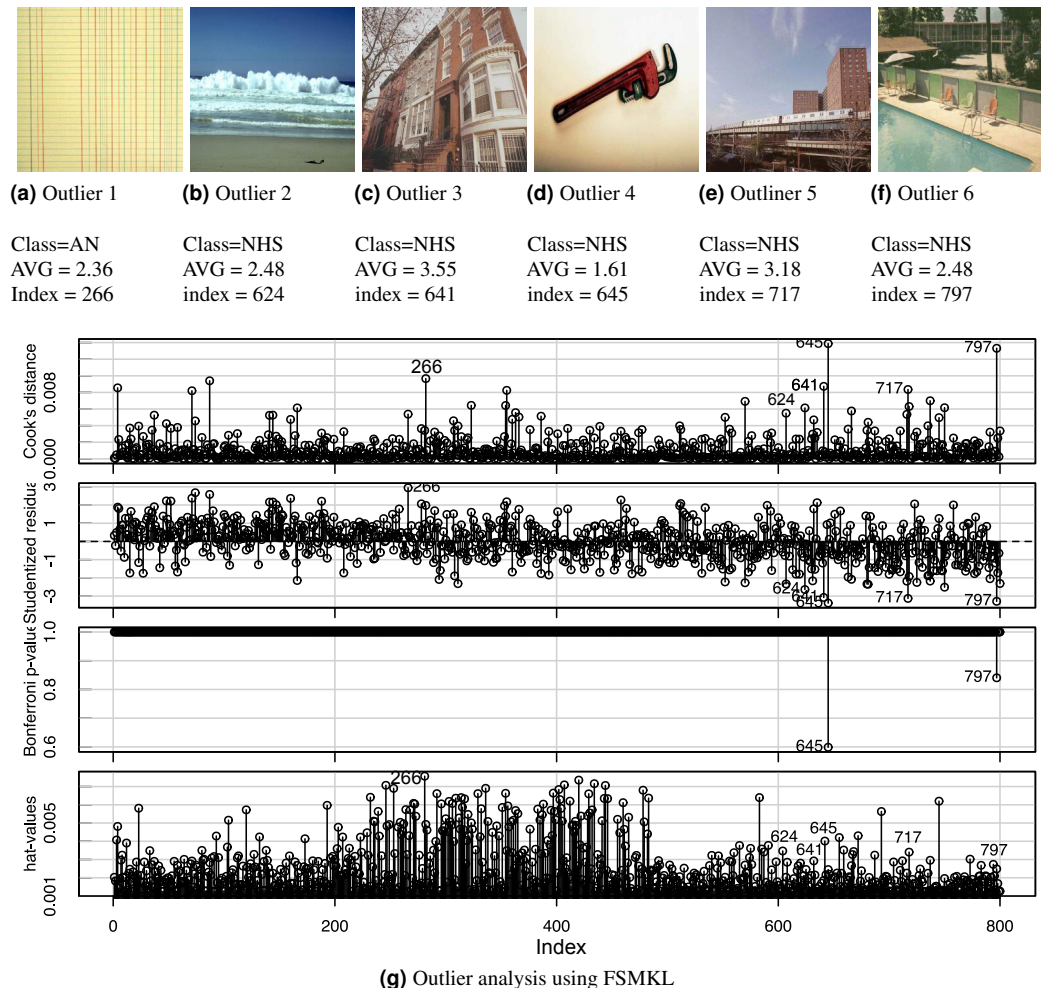
Figure 8. Six images were found, which were clearly outliers (a-f). An outlier analysis was performed, paying attention to: g.1) the influence of each particular image on the final model according to Cook's distance measure, g.2) the Studentized residuals (residuals divided by their estimated standard deviation), g.3) the p-values of the outlier test (Bonferroni correction) and g.4) the hat-values in order to identify influential observations

It was found that the other three initially proposed methods (LM, GLM and Lasso) increase the R-squared performance if the 50 experiments are run again without the outliers and according to a pairwise Wilcoxon test Wilcoxon (1945), statistically significant with p-value $< 8.08 \times 10^{-9}$ for LM, p-value $< 5.15 \times 10^{-13}$ for GLM and p-value $< 3.84 \times 10^{-8}$ for Lasso. However, these results are, once again, worse than the ones obtained by ENET, FSMKL and SVM-RFE.

Given these results, the authors considered at this point that this strange behavior may only be related to the ENET. As can be seen in Figure 9a, there is a significant difference (p-value $< 2.2 \times 10^{-16}$) between the results obtained by ENET and FSMKL with and without outliers according to a Wilcoxon test. The authors also checked the significance of the difference between ENET, FSMKL and SVM-RFE without outliers using a Friedman test with the Iman-Davenport extension, and our results showed that, with a very high level of confidence, FSMKL is significantly better than the other with a p-value $< 1.92 \times 10^{-41}$. Finally, in order to ensure the power and validity of our results, the contrast estimation is shown based on medians for the best three models in Figure 9a using a heatmap. This estimation in non-parametric statistics is used to compute the real differences between the algorithms García et al. (2010); Doksum (1967).

On the other hand, Figure 9b shows the distribution of the residuals achieved in the correlation results
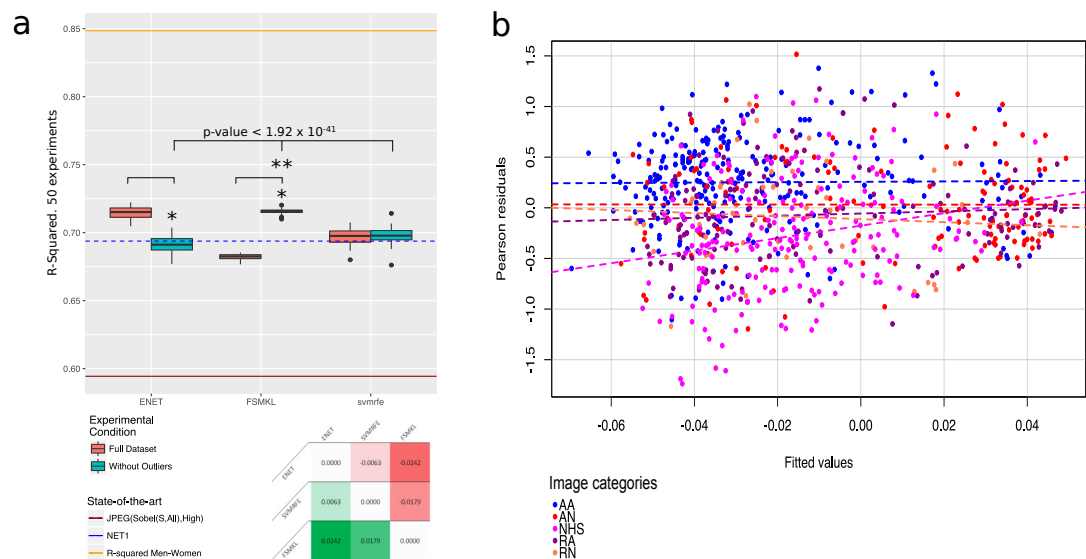
Figure 9. Outliers plots: a) boxplot with the results of the best three methods (ENET, FSMKL and SVM-RFE) with and without the outliers. * Statistically significant difference with a p-value $< 2.2 \times 10^{-16}$ according to a pairwise Wilcoxon test. ** Statistically significant difference according to a non-parametric Friedman test with Iman and Davenport correction with a p-value $< 1.92 \times 10^{-41}$. At the bottom of the panel, a median-based contrast estimation heatmap. In panel b) the residues are plotted by coloring each of them according to the category to which the image belongs.

322 of FSMKL. In order to clarify the belonging of each residual to a particular group (please refer to section
323), each dot was plotted in a different color. It should be noted that the slope of the AA, AM, RA and RN
324 groups was very similar and almost parallel between them. However, the slope of the NHS group was
325 steep and had a very different angle with respect to any other slope, in fact, this line intersected clearly
326 all the others. This reveals the correlation between the first four groups and also the low or nonexistent
327 correlation between those groups and NHS. The same was highlighted by Machado et al. Machado et al.
328 (2015), where the authors stated that the content of the NHS group was largely cut off from the rest of the
329 groups. In view of the FSMKL results, the authors of the current study agree with them.

DISCUSSION

331 Previous studies, such as Palmer et al. Palmer et al. (2013); Palmer and Schloss (2010), have emphasized
332 the color importance in many visual fields. The results presented in this paper show that certain channels
333 of the HSV color model, in this case Saturation and Value, may be predominant in this type of problems.
334 Machado et al. (2015) initially showed the validity of different features identified by three criteria: family
335 or method, color channel and edge detection filters. The conducted experiments showed that, for example,
336 in the case of JPEG- and Fractal-related features, all combinations of the S and V channels, with or
337 without the Canny filter, provided the necessary information to obtain the best possible regression model.
338 It was also found that the features related to the different methods for calculation of the Zipf's Law
339 complemented satisfactorily those already mentioned in the same conditions of color channels and edge
340 detection filters.
341   The main objective of this work is the study of other computational methods to identify alternatives to
342 ANN already used to solve this problem. Up to date, the best results obtained was 0.6938 in terms of
343 R-Squared using an input set of 329 features. From the studied methods, at least 3 offered similar results
344 in terms of correlation coefficients using a significantly reduced number of metrics, FSMKL standing
345 out with 22, which were repeated continuously when performing 50 independent experiments and with
346 an RMSE error with small variability. This stability of results allowed us to identify this method as the

most reliable for the studied problem. FSMKL is an integrative kernel-based technique, since, in addition to identifying the metrics best adapted to the problem, it also looks for the relationships between them that best fit the objective of the experiment, making the most of the complementarity of the features to increase the obtained score. That is, it makes the most of the correlation between the variables to use them in a complementary way and to add more knowledge to the learning process. More specifically, it integrates information from different image descriptors and is able to weigh the degree of similarity of each subset, allowing in this way to select the features or feature sets of each of the groups that jointly obtain the best results.

As part of this study, the possible outliers of the most stable model were analyzed, according to which 5 images belonging to the NHS set and 1 image of the AN set were identified. These outliers may be due not only to their visual nature, but also to the descriptors used, since reliably detecting edges of both groups is a difficult task Machado et al. (2015). It would be interesting to study other sets of descriptors not related to complexity estimators or edge detectors in order to check this issue.

Taking into account the residual values of the resulting model, it was also possible to identify the difference between the NHS group and the rest. This difference is comprehensible since it consists of photographs, while the other four include paintings or cliparts, in which effects such as brushstroke are susceptible to identification by methods of border detection, whereas cliparts tend to be images less complex than a painting at computational level. The obtained results are in accordance with what was initially proposed in the literature as possible, and can be demonstrated.

Although it is not part of this study, it would be interesting to observe these computational models individually with each group. However, the few available examples may distort the results because of overtraining or overfitting problems, such as the one experienced with the ENET model in this paper, since the number of examples would be lower than the number of features. It would be necessary in this case to add an external cross-validation process, in order to avoid possible selection bias of the results during the process of selecting the best technique parametersAmbroise and McLachlan (2002).

On the other hand, if one could observe the behavior after having removed the NHS group, already identified as conflictive, so that there would be 600 images divided into 4 groups, which may be compared pairwise according to different criteria.

CONCLUSIONS

This study proposed the application and validation of a set of different computational Machine Learning models for the evaluation of visual complexity. A set of 800 images were used (there were five different groups of images) and more than 300 features were calculated based on image compression error and Zipf's Law over three color channels.

Six state-of-the-art Machine Learning regression algorithms were compared, with and without feature selection, and a final statistical analysis step was performed in order to evaluate the results and select the most promising one.

The novel FSMKL was chosen, and from an exhaustive outlier analysis it was found that extreme images were present in the dataset and modified the regression value. Furthermore, it was proven that one of the groups containing 200 photographs of natural and human-made scenes (NHS) have a very low correlation with the other groups, as previously suggested by Machado et al. Machado et al. (2015).

Our results are of relevance, as they outperformed all the previous published works and are in accordance with the psychological findings of the human conception of visual complexity.

ACKNOWLEDGMENTS

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

REFERENCES

Aksentijevic, A. and Gibson, K. (2012). Psychological complexity and the cost of information processing. *Theory and Psychology*, 22(5):572–590.

Alario, F.-X. and Ferrand, L. (1999). A set of 400 pictures standardized for french: Norms for name agreement, image agreement, familiarity, visual complexity, image variability, and age of acquisition. *Behavior Research Methods, Instruments, & Computers*, 31(3):531–552.

Alfons, A. (2012). *cvTools: Cross-validation tools for regression models*. R package version 0.3.2.

Ambroise, C. and McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences*, 99(10):6562–6566.

Amirshahi, S. A., Hayn-Leichsenring, G. U., Denzler, J., and Redies, C. (2014). *JenAesthetics Subjective Dataset: Analyzing Paintings by Subjective Scores*, volume 8925 of *Lecture Notes in Computer Science*, pages 3–19. Springer.

Analytics, R. and Weston, S. (2015). *doMC: Foreach Parallel Adaptor for 'parallel'*. R package version 1.3.4.

Bhlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Publishing Company, Incorporated, 1st edition.

Birkhoff, G. (1933). *Aesthetic Measure*. Harvard University Press.

Blum, A. L. and Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1):245 – 271.

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. *Knowledge and Information Systems*, 34(3):483–519.

Canny, J. (1986). A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698.

Cela-Conde, C. J., Ayala, F. J., Munar, E., Maestú, F., Nadal, M., Capó, M. A., del Río, D., López-Ibor, J. J., Ortiz, T., Mirasso, C., and Marty, G. (2009). Sex-related similarities and differences in the neural correlates of beauty. *Proceedings of the National Academy of Sciences*, 106(10):3847–3852.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

Cycowicz, Y. M., Friedman, D., Rothstein, M., and Snodgrass, J. G. (1997). Picture naming by young children: Norms for name agreement, familiarity, and visual complexity. *Journal of experimental child psychology*, 65(2):171–237.

Dash, M. and Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis*, 1(1):131 – 156.

Do, K.-A., Quin, S., and Vannucci, M. (2013). *Advances in statistical bioinformatics: models and integrative inference for high-throughput data*. Cambridge University Press.

Doksum, K. (1967). Robust procedures for some linear models with one observation per cell. *Annals of Mathematical Statistics*, 38:878–883.

Donderi, D. C. (2006). Visual complexity: a review. *Psychological bulletin*, 132(1):73.

Eysenck, H. J. and Castle, M. (1971a). Comparative study of artists and nonartists on the maitland graves design judgment test. *Journal of Applied Psychology*, 55(4):389–392.

Eysenck, H. J. and Castle, M. (1971b). Comparative study of artists and nonartists on the maitland graves design judgment test. *Journal of Applied Psychology*, 55(4):389–392.

Fernandez-Lozano, C., Gestal, M., Munteanu, C. R., Dorado, J., and Pazos, A. (2016a). A methodology for the design of experiments in computational intelligence with multiple regression models. *PeerJ*.

Fernandez-Lozano, C., Seoane, J. A., Gestal, M., Gaunt, T. R., Dorado, J., and Campbell, C. (2015). Texture classification using feature selection and kernel-based techniques. *Soft Computing*, 19(9):2469–2480.

Fernandez-Lozano, C., Seoane, J. A., Gestal, M., Gaunt, T. R., Dorado, J., Pazos, A., and Campbell, C. (2016b). Texture analysis in gel electrophoresis images using an integrative kernel-based approach. *Scientific reports*, 6:19256.

Forsythe, A., Nadal, M., Sheehy, N., Cela-Conde, C. J., and Sawey, M. (2011). Predicting beauty: Fractal dimension and visual complexity in art. *British Journal of Psychology*, 102(1):49–70.

Fox, J. and Weisberg, S. (2011). *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, second edition.

Friedenberg, J. and Liby, B. (2016). Perceived beauty of random texture patterns: A preference for

complexity. *Acta Psychologica*, 168(Supplement C):41 – 49.

from Jed Wing, M. K. C., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., and Candan., C. (2016). *caret: Classification and Regression Training*. R package version 6.0-68.

García, S., Fernández, A., Luengo, J., and Herrera, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences*, 180(10):2044 – 2064. Special Issue on Intelligent Distributed Information Systems.

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

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1):389–422.

Hall, M. A. and Smith, L. A. (1999). Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper. In *Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference*, pages 235–239. AAAI Press.

Hayn-Leichsenring, G. U., Lehmann, T., and Redies, C. (2017). Subjective ratings of beauty and aesthetics: Correlations with statistical image properties in western oil paintings. *i-Perception*, 8(3):2041669517715474.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580.

Hocking, R. R. (1976). A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics*, 32(1):1–49.

Jain, A. and Zongker, D. (1997). Feature selection: evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):153–158.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artif. Intell.*, 97(1-2):273–324.

Lang, P. J., Bradley, M. M., and Cuthbert, B. N. (2008). International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical report, University of Florida, Gainesville, FL.

Lawrence, S., Giles, C. L., and Tsoi, A. C. (1997). Lessons in neural network training: Overfitting may be harder than expected. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence*, AAAI'97/IAAI'97, pages 540–545. AAAI Press.

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

Lempel, A. and Ziv, J. (2006). On the complexity of finite sequences. *IEEE Trans. Inf. Theor.*, 22(1):75–81.

Lyssenko, N., Redies, C., and Hayn-Leichsenring, G. U. (2016). Evaluating abstract art: Relation between term usage, subjective ratings, image properties and personality traits. *Frontiers in psychology*, 7:973.

Machado, P., Romero, J., Nadal, M., Santos, A., Correia, J., and Carballal, A. (2015). Computerized measures of visual complexity. *Acta Psychologica*, 160:43 – 57.

Maldonado, S., Weber, R., and Basak, J. (2011). Simultaneous feature selection and classification using kernel-penalized support vector machines. *Information Sciences*, 181(1):115 – 128.

Marin, M. M. and Leder, H. (2013). Examining complexity across domains: relating subjective and objective measures of affective environmental scenes, paintings and music. *PloS one*, 8(8):e72412.

Martinez, A. and Benavente, R. (1998). The ar face database, cvc, univ. autonoma barcelona, barcelona. Technical report, Spain, Technical Report 24.

Menden, M. P., Wang, D., Guan, Y., Mason, M., Szalai, B., Bulusu, K. C., Yu, T., Kang, J., Jeon, M., Wolfinger, R., Nguyen, T., Zaslavskiy, M., , Jang, I. S., Ghazoui, Z., Ahsen, M. E., Vogel, R., Chaibub Neto, E., Norman, T., Tang, E. K., Garnett, M. J., Di Veroli, G., Fawell, S., Stolovitzky, G., Guinney, J., Dry, J. R., and Saez-Rodriguez, J. (2017). Community assessment of cancer drug combination screens identifies strategies for synergy prediction. *bioRxiv*.

Palmer, S. E. and Schloss, K. B. (2010). An ecological valence theory of human color preference. *Proceedings of the National Academy of Sciences*, 107(19):8877–8882.

507 Palmer, S. E., Schloss, K. B., and Sammartino, J. (2013). Visual aesthetics and human preference. *Annual*
508 *review of psychology*, 64:77–107.

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

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

511 R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for
512 Statistical Computing, Vienna, Austria.

513 R. Dennis Cook, S. W. (1997). Graphics for assessing the adequacy of regression models. *Journal of the*
514 *American Statistical Association*, 92(438):490–499.

515 Rakotomamonjy, A., Bach, F. R., Canu, S., and Grandvalet, Y. (2008). SimpleMKL. *Journal of Machine*
516 *Learning Research*, 9:2491–2521.

517 Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics.
518 *Bioinformatics*, 23(19):2507–2517.

519 Schettino, A., Keil, A., Porcu, E., and Müller, M. M. (2016). Shedding light on emotional perception: In-
520 teraction of brightness and semantic content in extrastriate visual cortex. *NeuroImage*, 133(Supplement
521 C):341 – 353.

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

524 Sobel, I. (1990). An isotropic 3 x 3 image gradient operator. *Machine Vision for Three-Dimensional*
525 *Scenes*, pages 376–379.

526 Street, N., Forsythe, A. M., Reilly, R., Taylor, R., and Helmy, M. S. (2016). A complex story: Universal
527 preference vs. individual differences shaping aesthetic response to fractals patterns. *Frontiers in human*
528 *neuroscience*, 10:213.

529 Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical*
530 *Society. Series B (Methodological)*, 58(1):267–288.

531 Tsiliki, G., Munteanu, C. R., Seoane, J. A., Fernandez-Lozano, C., Sarimveis, H., and Willighagen, E. L.
532 (2015). Rregrs: an r package for computer-aided model selection with multiple regression models.
533 *Journal of Cheminformatics*, 7(1):1–16.

534 Waldmann, P., Mészáros, G., Gredler, B., Fürst, C., and Sölkner, J. (2013). Evaluation of the lasso and the
535 elastic net in genome-wide association studies. *Frontiers in Genetics*, 4:270.

536 Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

537 Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1(6):pp. 80–83.

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

540 Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort: An Introduction to Human*
541 *Ecology*. Addison-Wesley.

542 Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the*
543 *Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.