Authorship and aesthetics experiments. Comparison of results between Human and Computational Systems

Luz Castro, Rebeca Perez, Antonino Santos, and Adrian Carballal

Department of Information and Communication Technologies
University of A Coruña, A Coruña, Spain
{maria.luz.castro,rebeca.perezf,nino,adrian.carballal}@udc.es

Abstract. This paper presents the results of two experiments comparing the functioning of a computational system and a group of humans when performing tasks related to art and aesthetics. The first experiment consists of the identification of a painting, while the second one uses the Maitland Graves's aesthetic appreciation test. The proposed system employs a series of metrics based on complexity estimators and low level features. These metrics feed a learning system using neural networks. The computational approach achieves similar results to those achieved by humans, thus suggesting that the system captures some of the artistic style and aesthetics features which are relevant to the experiments performed.

1 Introduction

Art may be considered to be innate to human beings. We have learnt to use our bodies and the tools around us from the beginning of recorded history, not just to communicate but also to express our artistic motivations.

The development of computing has inevitably led to the pursuit of generating systems which are capable, not just of supporting artists, but also of complementing them. Ada Lovelace, the daughter of the famous poet Lord Byron and arguably the first programmer in History, forecast in her writings the possibility of creating computers with artistic capabilities.

However, creating computer systems which can automatically perform artistic tasks is a complex and often controversial field. Even the definition of concepts such as art, beauty and aesthetics generate complex debates in areas such as Philosophy and Art Psychology. Moreover, making art has always been subject to a considerable subjective element where individuals and the emotional and social environments surrounding them have a direct impact.

In spite of these difficulties, there are several researchers who explore the creation of computational systems related to art and aesthetics. Some experiments aim at the achievement of systems which are capable of classifying and evaluating images without needing an interaction with users [6].

Following this line, it is necessary to have a component capable of "perceiving" a work of art and performing its classification/ordering/evaluation. In other

words, this component would carry out the image classification or ordering according to some aesthetic or artistic features. There are a great number of papers presenting experiments related to these types of systems. We should highlight the special issue of the Journal of Mathematics and the Arts [4], and some previous papers at this conference, as well as in other fields [1, 5, 7]. This system would possess a great application by itself when integrated in search engines, as a pedagogical application or as support to artistic researchers all over the world.

From our point of view, creating these systems is hugely relevant within the framework of research into computational aesthetics. This paper compares an artificial system and those of human beings with different artistic training. For this purpose, two tests were performed, based on the validation methodology presented in [12]: one consisting of identifying the authorship of works by three different painters and another one based on aesthetic evaluation and appreciation by means of the psychological test developed by Maitland Graves. Both, the results of comparing humans and a computational system and the worth of the proposed metrics may be significant in this context.

2 Experiments involving humans

This section presents the experimental design of both tests with humans.

2.1 Authorship experiments

How can humans acknowledge the authorship of a painting? What kind of features, colors or shapes should it have in order to be classified within the author's own style? What makes us say, when faced with a work of art; this is a "Mondrian" or a "Van Gogh"?

In order to start the experiment, we consider style as a system of shapes with significant qualities and expression through which the artist's personality becomes visible, as well as the perspective of a group framed in time. A given style consists mainly of a series of elements of form and motives interconnected by the so called "artist's expression". Nevertheless, the very creator of works of art is submitted to a huge subjective and contextual criterion which makes their work vary through time. Thus, although their characteristics remain constant, their works may be framed within several periods, touching upon different styles. This can be seen, for example, in the different works by Van Gogh, from the cool hues and realism of "The Potato Eaters" from 1885 to, for instance, the famous and well known "Starry Night" from 1889 (Figure 1). Those people who are not experts in art may well identify both works as belonging to two different authors, showing the great relevance of a well trained CAA for these types of problems.

For this reason, for the purpose of this experiment it was decided that we would work with paintings by Picasso, Kandinsky and Monet. The three of them show in their works paintings from different styles and periods which make their identification difficult. Nevertheless, they are recognizable enough by the general public so that they constitute a sound basis.

(a) The Potato Eaters (Realism) (b) The Starry Night (Postimpressionism)

Fig. 1: Example of paintings belonging to the same author (Van Gogh) but framed in different periods.

A total of 666 images by the already mentioned artists were selected, form various stages and styles, taken from the Internet, and distributed as follows: 212 images of Picasso works, 339 by Monet and 115 by Kandinsky. The experiment was carried out under controlled conditions within the University of A Coruña. Sixty-two humans took part, most of them university students between the ages of 18 and 25 (28 females and 34 males). Each subject was requested to evaluate a total of 30 works randomly chosen and distributed equally among the three authors. A high number of evaluations allow the elimination of false positives and avoiding the possibility of only evaluating the most recognizable works by the authors by chance, thus biasing results.

Every work was treated before the experiment in order to eliminate any kind of seal or signature by the author which would give hints to the subjects. Moreover, an application was designed covering the whole display on computers without Internet connection, so that each user was monitored and their answers recorded and processed through a binary code.

Subjects were handed a reference book acting as previous training before the performance of the identification test. The book contains 27 images by each author and an identifier ("A","B","C") instead of their names. These images were also anonymous. Subjects were constantly monitored in order to stop them from using support material and no external aid was provided.

Subjects were asked to answer a series of questions about themselves so as to allow a bias based on their sex, age and artistic experience, so as to compare if there is any difference in the percentage of right answers among the group with some kind of experience in the art field (to be called from now on "Art1") and those who did not ("Art0").

Next, an application displays randomly and subsequently each of the 30 images in the group. The user must only mark the type he/she considers that the work belongs to (Type A corresponds to Kandinsky, Type B to Monet and Type C to Picasso) and then click on the "next" button. Based on previous studies showing that the time during which the image is displayed does not influence

aesthetic preference significantly, no maximum time limit was set. Users know at all times how many images they have evaluated and how many are left, however, they are not allowed to return at any time. At the end of the test, they are shown the achieved result.

2.2 Aesthetics Appreciation

The second experiment is focused on Maitland Graves' psychological test. This test yields the capacity for acknowledging some basic principles of aesthetic nature defined by the author; such as unity, predominance, balance among elements, variety, continuity, symmetry, proportion and rhythm.

Users in our experiment were provided with a short description of the test goal and the procedure to be followed. Thus, each individual was shown 30 items comprising two or three designs which were similar to each other (Figure 2). These 30 designs were randomly selected from the 90 in the test. In each item, one of the designs corresponds to the already mentioned criteria, thus being a correct image, while the other one (or the other two) do not comply with one or more of these principles.

The average percentage of correct answers resulting from answering randomly to the test is 48.3%, due to the fact that some of the items were made up of three images.

In addition to the original work of Graves, which shows an average percentage of correct answers of 49.4, there are different works that show the results obtained with this test on different samples of humans [3]. Eysenck and Castle [2] obtained a 64.4% success in populations with artistic knowledge. The Portuguese Institute for Employment and Vocational Training [9] obtained a 61.87% success in the case of students of Fine Art degrees vs random Portuguese individuals.

Also there are studies in which instead of human populations have been using mathematical and computational models [9]. In this case, Romero et al. obtained a success rate of 64.9% using a heuristic approach and a 71.67% through artificial neural networks.

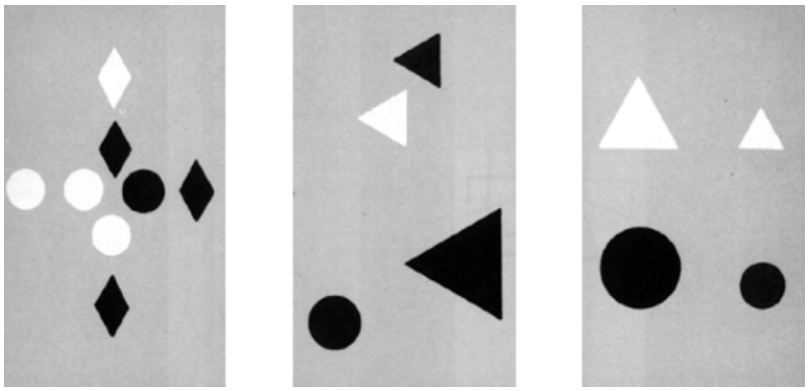
3 Developed System

The developed system is based on a feature extractor and a neural network.

The feature extractor shown in [9] has been available. The extractor is based on low level metrics, edge detection filters and complexity estimators inspired by [8]. To start with, the extractor standardizes images to a format of 128 x 128 pixels in order to avoid the different relations between width and height, thus facilitating the extraction process. Later on, the image is divided into three channels: hue, saturation and value (HSV). Four families of metrics are then extracted from the image: metrics based on compression error, Fractal Dimension based on Zipf distribution and statistics (mean and typical deviation).

In compression-based metrics, the ratio between the error generated by a compression method and the compression ratio is calculated (this calculation is

(a)



(b)

Fig. 2: Examples of images from the psychological test where humans tend to choose the wrong image.

made for JPEG and fractal compression, with three compression ratios in each case). Those metrics based on Zipf's Law use two values per channel, corresponding to the slope value and the correlation with the slope of the Zipf distribution of the value of the image pixels. For the value channel, the extractor also determines the Fractal Dimension of the original image and of the same one, after having applied three different types of Sobel filters. These filters allow a more accurate vision and identification of the image edges by approaching the gradient according to the image intensity. The Fractal Dimension was measured by means of using a similar technique to the one used by Taylor et al. [13], consisting of a simple method in which the image is turned into black and white, while the Fractal Dimension is estimated by means of the "box-counting" technique. This method allows cutting the image into small "boxes" and analyzing each one of them separately so as to obtain more satisfactory results. A Sobel filter

was used in order to obtain the Fractal Dimension of the edges. Later on, the Fractal Dimension of the resulting image was calculated in total.

With the purpose of determining the variation of the features which are inherent to the painting, the image was divided into five equally-sized regions: four squares and one big rectangle overlapped in the center. Later on, the metrics of each partition are calculated.

Using each of these sub-images provides information about features which are as relevant as symmetry and balance. This process entails a total of 246 measurements. The neural network is a backpropagation MLP with 246 neurons in the input layer, 12 in the hidden one and 3 in the output layer. Thirty independent repeats were carried out during the training phase for each of the neural networks with the goal of achieving statistically significant results. Training, test and validation sets were randomly created for each repeat, containing 70%, 10% and 20% of the patterns, which were later on accurately applied to each of the different architectures in the neural networks. The training of the neural networks ends when any of these criteria is met: 1500 training cycles or an average quadratic error in the training and test phases lower than 0.005. These parameters and the topology were empirically established in previous experiments of different research groups [10, 11].

In the aesthetic evaluation experiment, there were a very small number of patterns, which entailed the risk that the results were scarcely significant if the choice of training sets was not the right one. Therefore, a technique named on 20-fold cross-validation was carried out, which, in this case, yielded satisfactory results. Patterns were randomly distributed in 20 independent sets with a similar size among them (18 of them contain 5 patterns and two of them contain 4 patterns). Out of these 20 independent sets, 19 were used for the training, while only one of them was used for the test. Thus, none of the test patterns was included during the training, thus guaranteeing that the result has not been biased by a previous training. This process was repeated 20 times for every set, so that the results obtained comprise every possible case. An average of all the results obtained in each one of the sets is performed. The metrics corresponding to the Value channel of the image are used in this experiment (60 in total).

4 Results: Authorship test

For the authorship identification experiment of the three selected artists (Monet, Picasso and Kandinsky), the percentage of right answers was 81.82%. Tests were made with different combinations of metrics in order to evaluate their relevance. Figure 1 shows the results of the two best combinations of features and of the human groups. Net1 corresponds to all the metrics (246) and Net2 is similar to Net1 but without sub-images (41 metrics). A network trained with the metrics based on compression and those of mean and statistical deviation achieved a 78% rate of correct answers.

Kandinsky's images are those achieving the best classification, with a practically zero error percentage, while Picasso's images are the ones which get the

Table 1: Comparison of the results obtained by the human groups (Art0 and Art1) and the neural networks (Net1 and Net2) identifying authorship.

Approach	Description	Accuracy
Art0	38 individuals without art knowledge	83.42%
Art1	24 individuals with art knowledge	85.36%
Net1	264 metrics using all sub-images and all HSV channels	81.82%
Net2	41 metrics not using sub-images and using all HSV channel	78.00%

smallest percentage of right answers. These errors are usually found in paintings from styles or periods which are atypical in the painter, very distant from their best known works. For example, an observer could mistake the authorship of several Picasso works and those by Monet because of the treatment of form and color, as shown in Figure 3.

(a) Gabriele Münter Painting in Kallmünz (1903)

(b) In The Forest (1904)

Fig. 3: Examples of images by Kandinsky which observers tended to catalogue as either Monet's (a) or Picasso's (b)

As regards the experiment involving humans, the 62 individuals examined were divided into two sets: those who claimed not to possess any previous artistic experience (39 of them) (Art0) and those who claimed to have some artistic knowledge (23) (Art1).

The total rate of right answers is 84.43%. However, the rate of right answers in the group with some artistic experience is higher than that of the sample without any previous knowledge (85.36% vs. 83.42%). The images with the highest number of errors may be seen in Figure 4, 93.1% of errors (Fig. 4a) and 78.05% (Fig. 4b). Both are images by Kandinsky, attributed to Monet. The error percentage of Kandisnky's images is 71%, while the percentages in Picasso's and Monet's are around 90%.

(a) Doora Maar's portrait (1937)

(b) Crucifixion (1930)

Fig. 4: Examples of works which got the worst results in both neural networks authorship identification experiments.

5 Results: Aesthetic validation test

The aesthetic validation tests carried out consisted of Maitland Graves' psychological test: 90 items with the goal of identifying the best image from an aesthetic point of view. For this purpose, the ANN is simultaneously provided with metrics from both images.

The first architecture used with the ANN consists of 120 neurons (corresponding to the 60 features extracted from the Value channel of each image) in the input layer, 5 in the hidden layer, and 2 in the output layer. Its percentage of right answers was 66.33%.

With the aim of improving its evaluation and checking the importance of the metrics in the results elaboration, only 20 out of the 60 metrics proposed by the extractor were used in the next experiment: those corresponding to the

full image and to the rectangle superimposed at the center. Efficiency improved significantly, moving from a percentage of right answers of 66.33% to 70.41%.

The overall results from the 62 individuals correspond to only 46.2% of right answers, although those individuals with some previous artistic experience yielded better scores (42.56% vs. 52.32%). Table 2 shows a graph of the percentage of right answers by humans and computational systems. Systems A and B correspond to neural architectures 120-5-2 and 40-5-2.

Table 2: Comparison of the results obtained by human groups (Art0 and Art1) and neural networks (Net1 and Net2) in the aesthetic validation psychological test.

Approach	Description	Accuracy
Art0	38 individuals without art knowledge	42.56%
Art1	24 individuals with art knowledge	52.32%
Net1	60 metrics using all sub-images and Channel V	66.33%
Net2	20 metrics using one sub-image and Channel V	70.41%

6 Conclusions

The results of the comparison between a computational system with respect to a set of humans in carrying out tasks related to the art and aesthetics have shown as the system is able to recognize the studied paintings similarly to humans. In addition, the system also seems to be able to identify different aesthetic principles such as those used in the DJT better than the human population evaluated.

In the research, a computational system based on Artificial Neural Networks has been used with low-level metrics related to complexity, Fractal Dimension, Zipf, as well as typical deviation, average and the three color channels integrating the HSV (hue, saturation and luminosity). The experiment was carried out in two stages: one for identifying the authorship of a series of paintings and a validation test of aesthetic evaluation. Its performance has been compared with the results obtained by a set of 62 human, mostly university students.

Monet, Kandinsky and Picasso were chosen as painters to be classified for the authorship test, due to the wide dissemination of their works and the differences between their pictorial periods. The total number of works comprised 666 images by the painters (212 by Picasso, 339 by Monet and 115 by Kandinsky). Participants were asked to classify 30 random images into three main groups: Type "A", "B" and "C", each of them related to its author. They were not allowed to use any reference material, although they had been previously shown 27 images of each type, without providing any information about their authorship. The results achieved by humans have an average of right answers of 84.43%, while

those of the system achieved 81.82%. By using only compression error based metrics, together with the average and typical deviation, the result achieved is 78%. This result suggests that these metrics are the most significant ones in the proposed set.

Maitland Graves' "Design Judgement Test" was employed for the aesthetic evaluation tests. This consists of showing two or three images to the subject and asking him/her to indicate which one is the most correct. Only one of the images in the set complies with aesthetic composition criteria perfectly well, while the others do not comply with some of those principles. The results achieved by humans are 52.32% in the best-case scenario, vs. 74.49% by the system.

Nevertheless, it should be pointed out that the data is not comparable. The network has been previously trained with a set of images from the test itself, while participants in the experiment did not have any previous knowledge of it. Of course, it may be inferred that there is some degree of aesthetic sensitivity in human beings guiding them in their aesthetic judgement. However, this sensitivity should not always follow the principles defined by Maitland Graves in his test.

Anyhow, it may be deduced from the experiment results that the system, using just some metrics extracted from the image, has been capable of applying correctly some of the aesthetic principles defined by Maitland Graves in his test.

It may also be inferred that some of the metrics (or metric combinations) are capable of extracting or identifying the existence or inexistence of some of these principles.

Finally, the system has yielded similar results to those by humans in the authorship identification task from images. This suggests that the metrics used (specifically, those related to compression) allow a differentiation between different "styles" characteristic of each author, with results similar to those achieved by an average human being.

This work combines and compares the results of aesthetic appreciation from two different perspectives, the computational and the human, showing that in some cases the first might provide better results.

Acknowledgments

This research was partially funded by: Spanish Ministry for Science and Technology, research project TIN-2008-06562/TIN; Xunta de Galicia, research project XUGA-PGIDIT-10TIC105008PR.

References

1. Ekárt, A., Jo, A., Sharma, D., Chalakov, S.: Modelling the underlying principles of human aesthetic preference in evolutionary art. Journal of Mathematics and the Arts 6(2-3), 107–124 (2012)
2. Eysenck, H.J., Castle, M.: Comparative study of artists and nonartists on the maitland graves design judgment test. Journal of Applied Psychology 55(4), 389–392 (1971)

3. Götz, K., Götz, K.: The maitland graves design judgement test judged by 22. Perceptual and Motor Skills 39, 261–262 (1974)
4. Greenfield, G., Machado, P.: Special issue: Mathematical models used in aesthetic evaluation. Journal of Mathematics and the Arts 6(2-3) (2012)
5. den Heijer, E.: Evolving glitch art. In: Machado, P., McDermott, J., Carballal, A. (eds.) EvoMUSART. Lecture Notes in Computer Science, vol. 7834, pp. 109–120. Springer (2013)
6. Lewis, M.: Evolutionary visual art and design. In: Romero, J., Machado, P. (eds.) The Art of Artificial Evolution. pp. 3–37. Natural Computing Series, Springer (2008)
7. Li, Y., Hu, C., Minku, L., Zuo, H.: Learning aesthetic judgements in evolutionary art systems. Genetic Programming and Evolvable Machines 14(3), 315–337 (2013), http://dx.doi.org/10.1007/s10710-013-9188-7
8. Machado, P., Cardoso, A.: Computing aesthetics. In: Oliveira, F. (ed.) Proceedings of the XIVth Brazilian Symposium on Artificial Intelligence: Advances in Artificial Intelligence. LNCS, vol. 1515, pp. 219–229. Springer, Porto Alegre, Brazil (1998)
9. Machado, P., Romero, J., Manaris, B.: Experiments in computational aesthetics. In: Romero, J., Machado, P. (eds.) The Art of Artificial Evolution: A Handbook on Evolutionary Art and Music, pp. 381–415. Springer Berlin Heidelberg (2007)
10. Romero, J., Machado, P., Carballal, A., Osorio, O.: Aesthetic classification and sorting based on image compression. In: Chio, C.D., Brabazon, A., Caro, G.A.D., Drechsler, R., Farooq, M., Grahl, J., Greenfield, G., Prins, C., Romero, J., Squillero, G., Tarantino, E., Tettamanzi, A., Urquhart, N., Etaner-Uyar, A.S. (eds.) EvoApplications (2). Lecture Notes in Computer Science, vol. 6625, pp. 394–403. Springer (2011)
11. Romero, J., Machado, P., Carballal, A., Santos, A.: Using complexity estimates in aesthetic image classification. Journal of Mathematics and the Arts 6(2-3), 125–136 (2012)
12. Romero, J., Machado, P., Santos, A., Cardoso, A.: On the development of critics in evolutionary computation artists. In: Günther, R., et al. (eds.) Applications of Evolutionary Computing, EvoWorkshops 2003: EvoBIO, EvoCOMNET, EvoHOT, EvoIASP, EvoMUSART, EvoSTOC. LNCS, vol. 2611. Springer, Essex, UK (2003)
13. Taylor, R.P., Micolich, A.P., Jonas, D.: Fractal analysis of Pollock's drip paintings. Nature 399, 422 (Jun 1999)