Adaptive Critics for Evolutionary Artists

Penousal Machado[1], Juan Romero[2], María Luisa Santos[2],
Amílcar Cardoso[1], Bill Manaris[3]

[1] Centre for Informatics and Systems of the University of Coimbra, 3030 Coimbra, Portugal
{machado, amilcar}@dei.uc.pt
[2] Creative Computer Line, RNASA Lab. Faculty of Computer Science, University of Coruña, Spain
jj@udc.es, infmsa01@ucv.udc.es
[3] Computer Science Department, College of Charleston, Charleston, SC 29424, USA
manaris@cs.cofc.edu

Abstract. We focus on the development of artificial art critics. These systems analyze artworks, extracting relevant features, and produce an evaluation of the perceived pieces. The ability to perform aesthetic judgments is a desirable characteristic in an evolutionary artificial artist. As such, the inclusion of artificial art critics in these systems may improve their artistic abilities. We propose artificial art critics for the domains of music and visual arts, presenting a comprehensive set of experiments in author identification tasks. The experimental results show the viability and potential of our approach.

1 Introduction

The artistic process depends on the ability to perform aesthetic judgments, to be inspired by the works of other artists, and to act as a critic of one's own work. These factors depend on the artist's ability to see and listen. Modelling this capacity of the artist is an important step in the creation of an artificial artist. After all, an artist is also, and foremost, a viewer and listener. This view contrasts with the vast majority of the evolutionary computation systems for artwork generation (for a survey see, e.g., [1]), which tend to be completely blind/deaf to the outside world.

According to our view, the creation of a genuine evolutionary artificial artist requires the development of an Artificial Art Critic (AAC) – a system that is able to "perceive" an artwork, and perform an evaluation of the piece. The idea is to use the evaluations produced by the AAC to guide the evolutionary process.

In [2] we presented a general framework for the development of AACs. This framework consists of: an architecture, comprising a *feature extractor* and an *evaluator*; and a multi-stage validation methodology. The first stage includes identification tasks, such as the identification of the author or style of a given piece. This allows the objective, and meaningful, assessment of the AACs, providing a solid basis for their development. The later stages incorporate more subjective criteria, and include testing the AACs in a hybrid society of humans and artificial agents.

Following this set of ideas, we developed AACs for the musical and visual arts domains and conducted a broad set of experiments in the task of author identification.

2 System Description

The developed AACs are composed by two modules: a feature extractor and an evaluator. The feature extractor is static and domain specific. It is responsible for the perception of the artwork, generating as output a set of measurements that reflect relevant characteristics of the artwork. These measurements serve as input to the evaluator, which assesses the artwork according to a specific criterion defined by the user. The evaluator is an adaptive system, in our case implemented by means of an Artificial Neural Network. In the next sections we describe the modules used in the construction of the AACs. Namely, the musical and visual art feature extractors (section 2.1 and 2.2) and the adaptive evaluator (section 2.3).

2.1 Musical Feature Extractor

The musical feature extractor is similar to the one presented in [3]. It employs a series of Zipf's law [4] based metrics to extract features from music pieces encoded in MIDI format. Zipf distributions have been discovered in a wide range of phenomena including music. For instance, in [5] presents a study of 220 pieces of various music styles (baroque, classical, romantic, twelve-tone, jazz, rock, DNA strings, and aleatory music) discovering several Zipf distributions.

We use a total of 40 metrics, in addition to the number of notes of the piece. Each of the 40 metrics produces two real numbers:

1. The *slope* of the trendline of event frequencies plotted on a log-log, rank-frequency format; this number ranges from 0 to $-\infty$, with -1 denoting a Zipf distribution; and
2. The strength of the linear correlation, R^2, of the trendline; this ranges from 0 to 1, with 1 denoting a perfect fit.

The metrics used in the feature extractor can be divided into three types:

- Global metrics provide useful statistical information about the piece as a whole. There are seven metrics of this type: pitch, pitch-relative-to-octave, duration×pitch, duration×pitch-relative-to-octave, melodic interval, harmonic interval and melodic-harmonic interval.
- Structural metrics measure the equilibrium of higher orders of pitch change. Currently, we capture six orders of change. The first-order metric measures the equilibrium of changes between melodic intervals. The second-order metric measures the equilibrium of changes between first-order intervals, and so on.
- Fractal metrics measure the fractal dimension of each of the previous metrics. These metrics apply a given metric recursively at different levels of resolution within a piece. By successively subdividing a piece into parts, the lack of local balance can be exposed. Like the other metrics, fractal metrics produce a slope and a mean-square error value. The slope is equivalent to the fractal dimension of the given metric. The partitioning process stops when we reach phrases with less than five notes.

2.2 Visual Art Feature Extractor

The feature extractor presented herein is based on the notion that the complexity of an image is an important feature for the assessment of its aesthetical proprieties. This view is supported by a variety of studies (e.g. [6, 7])

In [8] the authors propose the use of complexity estimates to assess the aesthetical value of images, pointing the difference between complexity of the visual stimulus and complexity of the image perception task. The employed image and complexity estimates are based on the quality of jpeg and fractal image compression. A method for assigning aesthetic value according to these estimates is also presented. The approach was tested using a psychological test [9] designed to estimate the level in which an individual recognizes and reacts to basic principles of aesthetic order, achieving surprisingly good results. More recently, the same approach was used as part of an evolutionary art tool to assign fitness values to images and to filter images that are unquestionably bad [10]. Additionally, an image similarity metric based on these estimates is also presented.

The feature extractor proposed in this paper includes two types of complexity estimates: jpeg and fractal. To obtain these estimates we apply jpeg and fractal compression. The complexity of a given image is the ratio between the root mean square error resulting from its compression and the compression rate.

To better characterize the images, we vary the quality of the encoding by setting limits to the maximum error per pixel, and thus the amount of detail kept. This results in three complexity estimates for each compression technique.

After calculating these estimates for the whole image, the image is split into its Hue (H), Saturation (S) and Lightness (L) channels. We proceed by calculating the previously described estimates for each of the channels. Additionally we also calculate, for each channel: the average value; the standard deviation; the slope of the trendline of the Zipf distribution and root mean square error.

This process yields a total of 33 metrics, six for the whole image and nine for each of the three channels[1].

These global measurements can be misleading. For instance, the complexity of an image with three blank quadrants and a highly complex one can be similar to the complexity of an image with detail in all its quadrants. To get a better grasp of the distribution of these features by the different regions of the painting, we partition the image in five regions of the same size: the four quadrants, and an overlapping central rectangle. We apply to each region the previously described metrics. This yields a total of 198 measurements (33 for the entire image and 165 for the partitions). This number may seem too high but we prefer to use all the measurements for these initial experiments and to cut in a latter stage the ones that prove less meaningful.

2.3 Evaluator Module

The evaluator module is an adaptive system that uses the measurements made by the feature extractor as an input, and produces, as output, an evaluation of the artwork.

[1] Fractal image compression is not applied to the image as a whole since it would be redundant.

Being an adaptive module, the evaluator can adjust its behavior in order to perform different tasks. In this paper we focus in the task of author identification. Taking into account that we selected artists of different styles and movements to train and test our system, this task is, to some extent, also a style identification task.

From an architectural point of view, the adaptive evaluator consists of a feedforward ANN with one hidden layer. We use: standard backpropagation, with a learning rate of 0.2 and a momentum of 0, as learning function; the logistic function, as neuron activation function; and identity, as the output function. These settings remain unchanged throughout all the tests. The measurements made by the feature extractors are normalized between -1 and 1, before being fed to the network.

The number of measurements produced by the feature extractor determines the number of neurons of the input layer. Accordingly, the number of units in this layer is different for the two domains (music and visual arts). In the initial experiments in the musical domain we use an input layer composed by 81 neurons. In the visual arts domain, the input layer size is 198. In subsequent experiments we eliminate some of the inputs of the ANN, to assess the relevance of the different measurements. We conducted several preliminary experiments, varying the number of neurons of the hidden layer. In the experimental results presented in this paper we use hidden layers with 6 and 12 neurons, since these configurations gave the best overall results. The number of neurons of the output layer is equal to the number of authors considered for the test. In order to build, train and test the ANN we use SNNS[2].

3 Experimental Results

In order to train and test the ANNs we collected a significant amount of musical scores and artworks. We use a total of 741 scores, from a wide variety of music styles (e.g. prelude, fuga, toccata, mazurka, opera …). These scores were composed by five different authors, namely: Scarlatti (50 scores), Purcell (75), Bach (149), Chopin (291) and Debussy (176). We use a total of 802 different artworks, belonging to six different painters: 98 from Goya, 153 from Monet, 93 from Gauguin, 122 from Van Gogh, 81 from Kandinsky, and 255 from Picasso.

The training sets are constructed by randomly selecting a percentage (75% or 85%) of the available pieces. The test sets comprise the remaining ones.

3.1 Musical Domain

In this section we present some of the results achieved in a series of experiments in the musical domain. Due to space restrictions it is impossible to present all the performed experimentation.

[2] Stuttgart Neural Network Simulator (http://www-ra.informatik.uni-tuebingen.de/SNNS/).

Experiment 1. The aim of this experiment is to check the efficiency of the network in distinguishing between two authors (Scarlatti and Purcell) who belong to the same musical period (Baroque).

We used two different network architectures: 81-6-2 and 81-12-2. An output of (1,0) indicates a Scarlatti score, while (0,1) indicates a Purcell score. Table 1 summarizes the results of this experiment.

The experimental results show that the ANN is able to learn and generalize. The success rate on the test set varies between 100% and 90%, which corresponds to zero to four errors of identification (Table 1). As expected, using a training set with 85% of the scores yields better results. The results attained using 10000 learning cycles are globally worse than those achieved using a smaller number of cycles. This can be explained by the over-specialization of the network, which hinders its generalization abilities.

Table 1. Scarlatti vs. Purcell, two authors belonging to the same musical period

Train Set	Test Patterns	Architecture	Cycles	Errors	Success Rate	MSE	
						Train	Test
85%	19	81-6-2	10000	1	94.8%	0.00003	0.00003
			3000	0	100%	0.00021	0.00987
		81-12-2	10000	1	94.8%	0.00003	0.08408
			4000	1	94.8%	0.00020	0.06573
70%	37	81-6-2	10000	5	86.5%	0.00006	0.00023
			2000	4	90%	0.00023	0.17713
		81-12-2	10000	4	90%	0.00005	0.18011
			2000	3	92%	0.00026	0.12819

Experiment 2. The goal of this experiment is to distinguish between two composers that belong to different musical periods: Chopin (Romanticism) and Debussy (French Impressionism).

The architecture is identical to the one used in the previous experiment. In this case an output of (1,0) indicates a Chopin score, while (0,1) indicates a Debussy score. The results of this experiment are summarized in table 2.

Table 2. Chopin vs. Debussy, authors from different musical periods

Train Set	Test Patterns	Architecture	Cycles	Errors	Success Rate	MSE	
						Train	Test
85%	70	81-6-2	10000	0	100%	0.00001	0.00395
			3000	0	100%	0.00003	0.00358
		81-12-2	10000	0	100%	0.00001	0.00318
			3000	0	100%	0.00004	0.00205
70%	140	81-6-2	10000	1	99.3%	0.00001	0.01529
			4000	1	99.3%	0.00002	0.01356
		81-12-2	10000	1	99.3%	0.00001	0.01460
			5000	2	98.6%	0.00004	0.01848

As in the previous experiment, smaller training sets lead to a decrease of performance. However, in the current experiment the degradation of performance is minimal.

In this experiment the use of 10000 learning cycles appears to be adequate, there are no signs of over-fitting.

The results are clearly superior to the ones attained in experiment 1. There are two, possibly concurrent, explanations for the improvement of performance:

- The authors used in this experiment belong to different musical periods, which may make their discrimination easier;
- The number of musical scores available for the training (and testing) of the ANN is significantly larger, which may give a better coverage of the pieces created by the authors.

Experiment 3. The object of this experiment is the discrimination between three authors of three different musical periods: Baroque, Romanticism and French Impressionism. In this case the output layer is composed by three neurons. The output is interpreted as follows: (1,0,0) for a Purcell piece; (0,1,0) for Chopin; and (0,0,1) for Debussy.

Even though the composers belong to different schools, the success rates attained in this experiment are worse than the ones achieved in experiment 2. This is explained by the higher difficulty of the task, and also by the already noted difficulty in correctly identifying Purcell scores.

Table 3. Purcell vs. Chopin vs. Debussy, three authors of different musical periods

Train Set	Test Patterns	Architecture	Cycles	Errors	Success Rate	MSE	
						Train	Test
85%	82	81-6-3	10000	5	94%	0.00438	0.11285
			4000	6	92.7%	0.00452	0.11363
		81-12-3	10000	4	95.2%	0.00219	0.10078
			5000	4	95.2%	0.00005	0.10461
70%	162	81-6-3	10000	10	93.9%	0.00001	0.10437
			3000	10	93.9%	0.00257	0.11973
		81-12-3	10000	7	95.7%	0.00001	0.07450
			5000	8	95.1%	0.00002	0.07708

Experiment 4. The task in this experiment is discriminating between Bach compositions and works of other artists, namely: Scarlatti, Purcell, Chopin and Debussy. As before we use two output neurons, with an output of (1,0) indicating a Bach score, and an output of (0,1) indicating a composition made by other author.

The experimental results presented in table 4 show that this proved to be an easy task, which is undoubtedly due to the pronounced and highly recognizable style of this prolific composer.

Table 4. Experiment in order to recognize Bach from other composers

Train Set	Test Patterns	Architecture	Cycles	Errors	Success Rate	MSE	
						Train	Test
85%	106	81-6-2	10000	1	99.1%	0.00001	0.00997
			4000	0	100%	0.00316	0.00361
		81-12-2	10000	0	100%	0.00001	0.00413
			4000	0	100%	0.00315	0.00006
70%	217	81-6-2	10000	0	100%	0.00381	0.00246
			4000	1	99.6%	0.00382	0.00897
		81-12-2	10000	1	99.6%	0.00381	0.00288
			4000	1	99.6%	0.00382	0.00698

Experiment 5. The goal of this experiment is to check the efficiency of the network in distinguishing between the five considered composers. Accordingly the output layer is composed by five neurons. Since this was the more complete experiment performed we conducted a more exhaustive analysis, identifying the most relevant features for author recognition.

In order to assess the importance of each metric used in this experiment, we estimate the contribution of each metric for the output of the network. This is attained by calculating the sum of the absolute values of the weights between each input neuron and the neurons of the hidden layer. The reason for this procedure is the fact that the learning capacity of the biological neurons resides in their synapse, i.e. the intensity of their connections. In ANNs, the weights tag these connections with a value, which is related to the relevance of the associated neuron.

Table 5. Experiment with the five composers

Train Set	Test Patterns	Architecture	Cycles	Errors	Success Rate	MSE	
						Train	Test
85%	106	81-6-5	10000	6	94.4%	0.00005	0.07000
			4000	6	94.4%	0.00325	0.10905
		30-6-5	10000	7	93.4%	0.00319	0.12876
		15-6-5	10000	15	85.9%	0.01131	0.26105
		81-12-5	10000	6	94.4%	0.00313	0.11006
			4000	5	95.3%	0.00321	0.10201
		30-12-5	10000	10	90.6%	0.00348	0.12835
		15-12-5	10000	13	87.8%	0.00955	0.25258
70%	217	81-6-5	10000	11	95%	0.00386	0.09076
			4000	11	95%	0.00199	0.10651
		30-6-5	10000	14	93.6%	0.00386	0.10837
		15-6-5	10000	23	89.5%	0.02518	0.19658
		81-12-5	10000	14	93.6%	0.00194	0.14195
			4000	11	95%	0.00388	0.09459
		30-12-5	10000	11	95%	0.00195	0.08771
		15-12-5	10000	29	86.7%	0.02110	0.23455

According to this criterion for determining the importance of the features, we selected the 30 most relevant ones. We conducted several tests in which the input for the ANN was composed, only, by this set of features.

The experimental results (see table 5) indicate that these features are sufficient for the discrimination between the five authors. Therefore, we further reduced the set of input, conducting tests with the 15 most relevant features. The experimental results show a performance degradation. Using the 15 most relevant features yields an average success rate of 87.5%, using 30 yields an average success rate of 93.2%.

The analysis of the errors made by the ANN when identifying the test set patterns allows us to determine the most recognizable composers, and also the most challenging ones. As the results from experiment 4 led to believe, Bach was the most recognizable author. The most difficult author to recognize was Debussy. His works were often classified as scores of Chopin. Considering the remarkable influence of Chopin on composers such as Liszt, Wagner, Tchaikovsky, and especially Debussy, (melodic clashes, ambiguous chords, delayed or surprising cadences, remote or sliding modulations, unresolved dominant 7ths…) this does not come as a surprise.

An overall analysis of the results of the AAC in the different experiments reveals that they are coherent to the results that could be expected from a human. This reinforces the idea that the proposed feature extractor and evaluator are well suited to the task of identification.

3.2 Visual Arts Domain

Due to space restriction we only present the experimental results for the task of discrimination between six authors. The considered authors (with the exception of Gauguin and Van Gogh, which are usually considered Post-Impressionists) belong to different art movements. However, they all produced several works that are not within the style of the movement to which they are usually connected. Picasso, for instance, is usually associated with cubism; nevertheless he created a vast amount of non-cubist artworks. We faced two possible approaches for the collection of the artworks: using only the artworks that are more characteristic of a give painter; use a, hopefully, representative set of all the artworks produced by a given artist.

We chose the second approach. This choice was motivated by the following reasons: we needed a large set of training and test images; using only the most characteristic artworks may induce a bias, making the experiment artificially simple; the use of a more heterogeneous set of images increases the difficulty of the classification task.

The architecture of the ANN is similar to the used in the previous experiments. In this case the input layer is composed by 198 neurons. The number of neurons in the input layer was subsequently decreased to 148, 124, and finally to 99 neurons, in an attempt to identify the most relevant features.

To select the most relevant features we took into consideration the weights of the ANN using the configuration that yields the best results (198-12-6). The criterion used for the pruning of the input layer was the same as the one described in experiment 5.

Table 6. Experiment with the six painters

Train Set	Test Patterns	Architecture	Cycles	Errors	Success Rate	MSE	
						Train	Test
85%	120	198-6-6	10000	9	92.5%	0.00589	0.13879
		198-12-6	10000	4	96.7%	0.00299	0.09533
		148-12-6	10000	6	95%	0.00002	0.08357
		124-12-6	10000	6	95%	0.00297	0.06601
		99-12-6	10000	8	93.4%	0.00737	0.12137
70%	241	198-6-6	10000	22	90.9%	0.00537	0.15199
		198-12-6	10000	17	93%	0.00359	0.15364
		148-12-6	10000	14	94.2%	0.00182	0.12319
		124-12-6	10000	16	93.4%	0.00361	0.11943
		99-12-6	10000	14	94.2%	0.00360	0.09729

An analysis of the results presented in table 6, shows that the success rates achieved using the 198 measurements are similar to the ones attained using, only, the 99 most relevant ones. In fact, for some configurations the deletion of the inputs yields higher success rates. This indicates that those features are of small or no relevance. The use of non relevant features during the training stage usually hinders the generalization abilities of the ANNs.

An analysis of the errors made by the ANN in the classification of the test set instances, shows that the most recognizable painter is Gauguin, whereas Goya was the most difficult to identify. The total classification errors were distributed as follows: 4.3% while classifying Gauguin artworks; 8.5% on Van Gogh pieces; 9.4% on Monet; 10.1% on Picasso; 30.4% on Kandinsky; and 37.3% on Goya.

The difficulties in correctly identifying Goya's paintings were unexpected. An analysis of the artworks used in the training and testing of the ANN, indicates that the training instances do not provide a good coverage of the types of artwork produced by this author. Additionally, some of the paintings of this 18th century author were restored, which may pose further problems. The difficulties in the identification of Kandisnky's paintings can be explained by the heterogeneity of the Kandinsky's works included in the training and test set.

4 Conclusions and Future Work

In this paper we described the development of AACs for the domains of music and visual arts. The AACs follow a common architecture, and are composed by a feature extractor and an evaluator module. To validate our approach we tested their performance in several identification tasks.

The experimental results show that the proposed AACs are well-suited for these tasks, and that the extracted features are sufficient for the characterization of the pieces. Moreover, the analysis of the results allowed the identification of the most relevant features for the identification task, which is of key importance for the further development of our system.

The architecture used in the development of the AACs enables, the easy incorporation of additional features without damaging the performance, and the adaptation to other art domains.

Future research directions include: the further development of the feature extractor modules; assessing the performance of the system on a wider variety of tasks, including the aesthetic evaluation of the pieces; and the incorporation of the developed AACs in an evolutionary art system.

Research in the area of evolutionary artists is still on embryonic stage. The construction of AACs is an important step in the design of a true evolutionary artificial artist, and potentially in the better understanding of the artistic and creative process.

References

1. Johnson, C., Romero, J.: Genetic Algorithms in Visual Art and Music. In: Leonardo. MIT Press. Cambridge MA, Vol. 35, (2), (2002) 175-184
2. Romero, J., Machado, P., Santos, A., Cardoso, A.: On the Development of Critics in Evolutionary Computation Systems. In: Lecture Notes in Computer Science, Applications of Evolutionary Computing. LNCS 2611, Springer-Verlag, (2003) 559- 569
3. Manaris, B., Purewal, T. and McCormick, C.: Progress Towards Recognizing and Classifying Beautiful Music with Computers. In: Proceedings of EEE SoutheastCon, Columbia, SC. (2002) 52–57
4. Zipf, G.K.: Human Behavior and the Principle of Least Effort. New York: Hafner Publishing Company, (1949)
5. Manaris, B., Vaughan, D., Wagner, C., Romero, J., and Davis, R.: Evolutionary Music and the Zipf-Mandelbrot Law: Developing Fitness Functions for Pleasant Music. In: Lecture Notes in Computer Science, Applications of Evolutionary Computing. LNCS 2611, Springer-Verlag, (2003) 522–534
6. Arnheim, R.: Entropy and Art. University of California Press (1971)
7. Taylor, R. P., Micolich, A. P., and Jonas, D.: Fractal Analysis of Pollock's Drip Paintings. In: Nature (1999) 399–422
8. Machado, P., and Cardoso, A.: Computing Aesthetics. In: Oliveira, F., ed., Proceedings of the XIVth Brazilian Symposium on Artificial Intelligence SBIA'98. Porto Alegre, Brazil, Springer-Verlag, LNAI Series, (1998) 219-229
9. Graves, M.: Design Judgement Test Manual. The Psychological Corporation, New York, (1948).
10. Machado, P., and Cardoso, A.: All the truth about NEvAr. In: Bentley, P., Corne, D., eds., Applied Intelligence, Special issue on Creative Systems. Vol. 16, Nr. 2, Kluwer Academic Publishers (2002) 101–119