Artificial Music Critics

Juan Romero, PhD.
Creative Computer Line – RNASA Lab, Fac. of Computer Science, University of La Coruña, Spain.
e-mail: jj@udc.es

Penousal Machado, PhD.
Instituto Superior de Engenharia de Coimbra, Portugal
e-mail: machado@dei.uc.pt

Miss. María Luisa Santos Ares.
Creative Computer Line – RNASA Lab, Fac. of Computer Science, University of La Coruña, Spain.
e-mail: mhyso@hotmail.com

Abstract

This paper proposes a framework for the simplification of the development of Artificial Art Critics. We provide two basic elements: an architecture that consists of two main modules for the pre-processing and classification of an artwork, and a validation methodology that consists of several stages, such as the objective evaluation of an artwork (with targets like author or style identification) and a dynamic evaluation that implies the integration of the Artificial Art Critic into a multi-agent environment. We also present some experimental results concerning the first stage of the validation methodology. The results show the ability of the system to identify the author of a musical piece and its adaptive capacity to determine the relevant features of the musical piece.

1. Introduction

We believe that every artist, or in general every creator, has to act as a critic during the creation of a product, because he must judge whether or not it can be considered innovative or aesthetically pleasant. The creative process necessarily involves a phase of critique, parallel to the creative process but also an integrated part of it, that progresses at the same time as the creation itself.

The role of the critic as a key component of the creative process seems to have been forgotten by most of the artificial artwork generation systems developed over the last years. These systems ignore both their own generated artworks and external creations. In our opinion, the critic role of an artificial artist may not be neglected in a system that is really capable of acting as its human counterpart.

This paper describes a general framework for the development of an Artificial Art Critic (AAC) that consists of an architecture and a validation methodology, and it presents a series of experimental results concerning the application of this validation methodology.

The proposed architecture has two main modules. The *feature extractor* perceives the artworks and measures their characteristics. The *evaluator* generates an output based on these measurements and on additional feedback information. Depending on this feedback, the *evaluator* provides an objective assessment of the artwork (e.g. a style or author recognition task), or a "subjective" assessment (e.g. an evaluation task).

One of the main problems in the development of Artificial Artists or Artificial Critics is the validation of the fact that the system is behaving correctly or as expected: it is extremely complicated to specify what the system is expected to do. For an Artificial Artist, these expectations are very vague: we expect our system to create an innovative, aesthetically pleasant artwork.

The proposed multi-stage validation methodology intends to tackle this problem. Each stage of the methodology focuses on a different validation task: the first stage allows the objective and meaningful assessment of the system; the latter stages add more subjective criteria, and include testing the system in a hybrid society of humans and artificial agents.

This paper presents a series of experimental results in the first stage of the validation methodology.

2. Framework

We want to provide a basis for the validation and development of AMCs that allows the integration of contemporary critics and promotes collaboration between groups in the creation of AMCs. The overall framework is based on the following set of characteristics:

- Adaptability: an AMC should adapt to a changing environment. It must replicate a commonly accepted characteristic of human critics, namely evolution.

- Sociability: this can be seen as a concretion of the adaptability characteristic. AMCs will be integrated into a society with certain cultural and aesthetic trends, so they should be able to behave according to these trends. When performing in a hybrid environment – one that incorporates humans and artificial systems – the AAC must be validated by the society of artificial and human "agents", in the exact same way that human critics are validated in purely human societies [1].

- Independence of representation: the AMC should build its own internal representation of the artwork, shaping its assessment from the artwork itself; and it should only have access to the piece of art, not to any sort of higher level or external representation of the piece.

In addition, as the task of criticizing or expressing an opinion about an artwork is of the same nature, be it a music or whichever other kind of art, we have generalized the architecture in order to allow handling different artistic domains, adding a new characteristic to the list above:

- Generality: the AMC should be easily adaptable to different domains – becoming an Artificial Art Critic – so domain-specific tasks should be carried out by specialized modules, enabling the whole system to easily change from one domain to another.

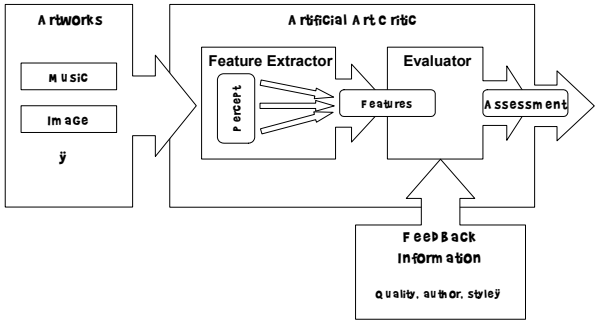

Figure 1. Outline of the proposed architecture

2.1 Architecture

If it intends to take into account the abovementioned characteristics, the architecture must allow the development of adaptive AMCs that will be easily adaptable to different domains (becoming a generic Artificial Art Critic, AAC), taking into account the particularities of each artistic domain. For instance, the way of dealing with music and with visual art is visibly different: whereas music follows a predetermined temporal sequence, the art viewer has a less constrained and more direct access to the piece of art. Hence the need to divide the system into various modules, providing specific ones that deal with domain particular tasks and allow the generality of the others.

Usually, artworks contain a huge amount of information. In visual art, for instance, even a relatively small picture can consume a large amount of memory. As one can infer from a state-of-the-art analysis of current adaptive systems (e.g. neural networks, genetic algorithms), such vast amounts of information cannot be handled reasonably. Some researchers try to tackle this problem by reducing the size of the artworks fed to the adaptive system (e.g. [2]). However, this approach implies an important loss of information and detail, and the experimental results are, typically, disappointing.

We believe that there is a more adequate approach, which consists of some kind of pre-processing of the artworks in order to extract relevant features, which can then be used as an input for the adaptive part of the system. This reduces the amount of information that has to be processed.

The proposed AAC architecture (roughly outlined in Fig. 1) includes two modules: the feature extractor and the evaluator. Each module has a concrete and different purpose. The feature extractor performs an analysis of the artwork and provides a set of relevant features to the evaluator. The evaluator makes an assessment of the artwork based on the previously extracted features.

The features extractor module performs two specific tasks: perception and analysis.

During the perception task the system builds some kind of internal representation of the artwork. Then, in the analysis task, this representation is analyzed and provides a set of relevant measurements. While this partition between perception and analysis is mainly conceptual, the idea behind it is that, in a first stage, the features extractor acquires

information about domain-specific parameters which are then analyzed.

The internal representation is not constrained, nor are the techniques used on the feature extractor.

The evaluator module is an adaptive system that accepts, as an input, the features measurements carried out by the previous module, and processes this input so as to obtain, as an output, an assessment of the artwork.

The evaluator module must adapt to different tasks according to the feedback information provided. Depending on the task, this feedback can be an indication of the desired answer or an evaluation of the performance of the AAC, which must adjust its behaviour in order to maximize the performance. Also, the adaptive evaluator module provides information about which characteristics or measurements are relevant in the assessment of an artwork. The weights of an ANN, for example, give an indication of which characteristics are more significant when criticizing a piece. It is also possible to test the evaluator with different sets of features so as to find the minimum set needed for a particular evaluation task.

With this architecture, the search for new relevant features and evaluations remains independent: a system can include a set of different feature extractors and evaluators from different authors in order to test which combination of extractor and evaluator is preferable to characterize a piece of art.

We shall now present a validation methodology that allows us to test the developed AAC.

2.2 Validation Methodology

The validation of our AAC presents two main difficulties: the subjectivity involved in the evaluation of artworks, and the fact that large training sets are needed to train the evaluator module (hundreds of man-evaluated artworks).

The answer to these complications is the use of a multi-stage validation methodology. The AAC is presented with a different task in each level, starting with tasks in which the correctness of the ACCs output can be objectively determined and which do not require a training set of human evaluated artworks. We then move on to tasks that require more subjectivity and complexity. The response of the AAC is supposed to be static in the first levels; in the last levels, the AAC must adapt to the environment and change its evaluation over time according to the surrounding context.

At this time, the validation is divided into three levels: Identification, Static Evaluation and Dynamic Evaluation.

In the Identification level, the AAC recognizes the style or author of a given artwork.

During the Author Identification task, the AAC is presented with several artworks by different authors. Its task is to determine the author of each piece. The evaluator module can be trained by giving it feedback information that indicates the correct answer. This validation is easy to perform, the compilation of training instances is simple, and the test is absolutely objective. The main difficulty involved in this level of testing is the construction of representative training and test sets.

This validation step is limited in scope but it is useful in determining the capabilities of the feature extractor module. A failure during this test may indicate that the set of extracted features does not suffice to discriminate between authors: this prevents us from moving on to a more complex task, bound to fail due to the lack of meaningful information. The analysis of the features used by the evaluator to determine the correct author can help determine the relative importance of each of the extracted features. In fact, it is possible to perform specific tests to determine the predictive power of each measurement or set of measurements.

During the Style Identification task, the AAC must identify the style of an artwork. This type of validation allows the testing of AACs that may be used in a wide variety of tasks, such as image and music retrieval (and allowing, for instance, style-based searches).

It may be more difficult to discriminate between artists of the same school than to distinguish styles that are radically different. However, discriminating between artists that have characteristic signatures (in the sense used by Cope [3]) is easier than discriminating between closely related styles, so the difficulty of these tasks depends on the chosen artists and styles.

During the analysis of the experimental results, it is important to take into account what is reasonable to expect. For instance, if the testing set includes atypical artworks, the AAC will most likely fail. This does not necessarily indicate a flaw of the feature extractor or evaluator, but simply the fact that the artwork is atypical.

The Static Evaluation is the second level of validation. In this case, the AAC must determine the aesthetic value of a series of artworks previously evaluated by humans. To this effect, it needs a representative database of consistently evaluated artworks, whose construction is one of the major difficulties in the performance of this test.

The training of the AAC requires positive and negative examples, and, paradoxically, it is quite difficult to obtain a representative set of this type of examples.

The use of complexity appraisers in the feature extractor module may prove useful to rule out this type of items: representative samples of items that do not even meet the necessary requirements to be considered a piece, such as images in which the pixels are totally uncorrelated, and, as such, are nothing more than noise. The relation between complexity and aesthetic value has been pointed out by several authors (see, e.g., [4]); complexity appraisers have successfully been used as a way to filter images that do not meet the necessary pre-requirements to be considered artworks [5].

We could use a generative art tool to create the training set. This would yield a relatively high number of pieces in a reasonable amount of time. However, the set would only be representative of the pieces that are typically created by that generative art tool. Moreover, the degree of correlation between the created pieces may be high, making the task of the AAC artificially easy.

Another option would be to diminish the scope of application of the AAC: create an AAC that is able to assess the aesthetic quality within a well-defined style. This results in a validation step that is somewhat closer to the task of “Style Identification”, and as such less subjective. The difference is that the AAC is assessing the distance to a given style instead of trying to discriminate between styles.

The analysis of the experimental results can be challenging: one needs to make sure that the AAC is performing the expected task and not exploiting some flaw of the training set.

This type of problem is detected using the trained AAC to assign fitness to the pieces generated by an evolutionary art tool, and thus guide the evolution process. Evolutionary algorithms are especially good at exploiting holes in the fitness evaluation (see, e.g., [6]). Therefore, one can check if the evolutionary algorithm is able to generate abnormal pieces, which are highly valued by the AAC in spite of their poor quality.

The Static Evaluation step poses many difficulties, both in the construction of the test and in the analysis of the experimental results. It is, however, necessary in order to assess an AAC.

The last step in the methodology is the Dynamic Evaluation. The value of an artwork depends on its surrounding cultural context (or contexts). As such, the AAC must be aware of this context, and be able to adapt its assessment to changes in the surrounding environment.

To perform this validation, the "Hybrid Society" (HS) model is proposed. HS is a paradigm similar to Artificial Life, but with human "agents" at the same level as the artificial ones. HS explores the creation of egalitarian societies populated by humans and artificial beings in artistic (or other social) domains; as such, HS is adequate to validate the AAC in a natural and dynamic way. In the Dynamic Evaluation step, the success of the AAC depends on the appraisal of its judgments by the other members of the society. This type of test introduces a new social and dynamic dimension to the validation, since the value of an artwork varies over time, and depends on the agents that compose the society.

This validation level presents the problem of the need to incorporate humans in the experimentation. The experiments are difficult to plan and organize, and strong time limitations exist. Moreover, the adaptation capacity of the critics must be high in order to adapt to a dynamic and complex environment. In spite of the inherent difficulties, these critics can be valuable and easy to integrate in the "information society" as assistants of users or as part of general composers.

It is possible to assess the performance of the feature extractor and evaluator module independently, since the output of the feature extractor (in conjunction with the feedback information), can be seen as a training instance to the evaluator, and only in the first two levels of validation.

At the third level this is no longer possible since the feedback information does not directly reflect the quality of the artworks, but only an appraisal of the AAC actions by society, which changes dynamically in time.

The validation methodology presented here tries to find a compromise between automated and human-like validation.

3. Experimental Results

The experiments related to the development of an AAC are focused on the musical field. They distinguish the authors by means of their works and represent the first validation level of an artificial critic.

We have used 741 scores with an ample variety of music styles (prelude, fuga, toccata, mazurka, opera…) composed by a total of 5 different musicians: Scarlatti (50), Purcell (75), Bach (149), Chopin (291) and Debussy (176).

The system consists of two modules: the static feature extractor, presented by Manaris et al. (2003) [7], is based on a set of musical metrics and on the distribution of Zipf in order to obtain a series of values that represent each theme. The metrics surge from musical attributes (pitch, duration, melodic intervals, harmonic intervals,...) and are divided into simples (simple note and interval) and fractals (fractal note and interval). The Zipf distribution of each metric generates two numbers: the slope and the square average error (R2), which indicates the adjustment of the trendline to the value. This means that the output of the feature extractor is a set of numbers, two for each metric. We have used a total of 81 values: 40 metrics and the corresponding notes.

The adaptive evaluator consists of a feedforward ANN. We use the Backpropagation learning function, with a learning rate of 0.2 and a moment 0; the neuron activation function is the logistic function, whereas the output function is the identity function. Where the topology is concerned, the input layer consists of 81 units or neurons that correspond to the 81 values obtained in the previous stage. These values are normalized between -1 and 1 for their presentation to the network. We consistently used one single hidden layer, whose amount of units is variable (some tests have 6 units, some 12) so as to observe their influence on the final results. The output layer also varies according to the analysis that is being carried out (discrimination between two authors, three, five…). We use the SNNS application to construct, train and test the network. And in order to optimize the learning of the different patterns, we have presented them in a different order within the different cycles.

We use a high percentage (85%) of patterns in the training set (the test set consists of the remaining patterns), and two ANN architectures: one with 6 neurons in the hidden layer, the other with 12 neurons. Within each architecture we carry out an exploration with 10.000 cycles (a cycle is a training unit in which the file patterns are presented once to the network); after observing the error graphic (MSE, Medium Square Error) and their results, we analyze a different amount of cycles, which is usually smaller.

3.1. Experiment 1: Scarlatti vs. Purcell

The aim of this experiment is to check the efficiency of the network in distinguishing between two authors who belong to the same musical period (Baroque), and both to their initial years, which is an additional difficulty.

Architecture 81-6-2:

81 neurons in the input for the different metrics, 6 neurons in the hidden layer and 2 neurons in the output layer with the following representation: (1,0) to indicate that it is a work by Scarlatti and (0,1) by Purcell (Figure 2).

After 10.000 learning cycles, and in a test set of 19 patterns, the ANN identified all the authors except one, which indicates a success rate of 94,8% . The MSE was 0.00003 in the training set, and 0.04109 in the test set. After 3000 cycles, the ANN identified the authors of all the pieces of the training and test sets (100%). The MSE was 0.00021 in the training set and 0.00987 in the test set.

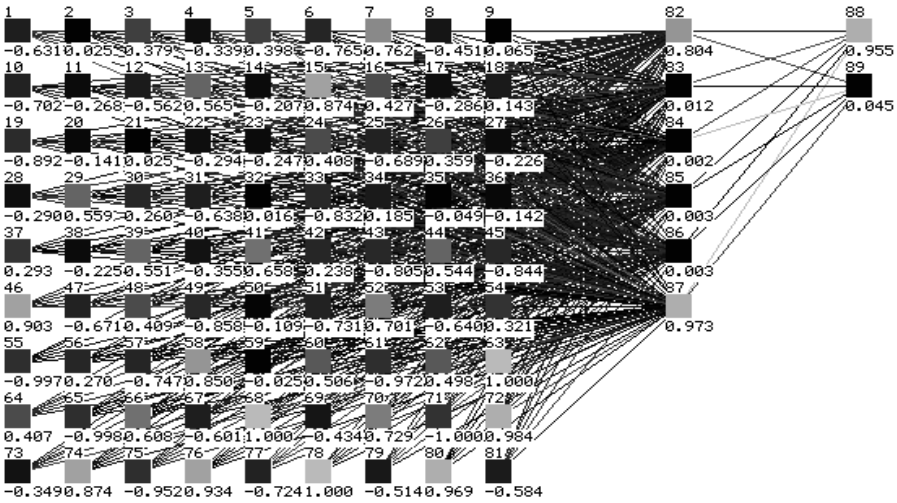

Figure 2. Architecture of the trained ANN

Architecture 81-12-2:

In this case, the number of units in the hidden layer amounts to 12. After 10.000 cycles, and as in the previous case, the ANN identified all the authors except one (94,8%). The MSE was 0.00003 in the training set and 0.08408 in the test set. After 4000 cycles (this amount allows us to see the evolution of the test set; in cycle 3000 it reaches its lowest value, 0.03353), we find the same situation, with a success rate of 94,8%. The MSE was 0.00020 in the training set and 0.06573 in the test set.

3.2. Experiment 2: Scarlatti vs. Purcell vs. Bach vs. Chopin vs. Debussy

Our aim is to check the efficiency of the network in distinguishing between these five composers. To this effect, we have carried out a more exhaustive analysis by identifying the most relevant features for the author recognition.

The results are as following:

Architecture 81-6-5:

In this case, the output layer contains 5 neurons and the activation of each neuron corresponds to a different author: (1,0,0,0,0) to identify Chopin, (0,1,0,0,0) for Bach, (0,0,1,0,0) for Scarlatti, (0,0,0,1,0) for Purcell, and (0,0,0,0,1) for Debussy.

After 10.000 learning cycles, the ANN made 6 mistakes in a test set of 106 patterns, which indicates a success rate of 94,4%. The training MSE was 0.00005 and the test MSE 0.07000. After 4000 cycles, we found 6 errors, as was the case with 10.000 cycles, which indicates the same error percentage. The training MSE was 0.00325 and the test MSE 0.10905.

Architecture 30-6-5:

In order to detect the importance of the metrics involved in this experiment, we checked the contribution of each metric to the network through the sum, in absolute terms, of the weights between the input units and the successor neurons (in this case, those of the hidden layer). This means that we added up the weights that connect neuron 1 with all the neurons of the hidden layer, neuron 2 with all the neurons of the hidden layer… and so forth until 81. Among all the obtained results, we chose the 30 heaviest weights, since they contribute more to the final result of the network.

During the training of the ANN with 30 neurons in the input layer (the rest of the architecture remains unchanged) and during 10.000 cycles, the results were the following: we observed 7 errors among the 106 patterns of the test file, which means that the ANN reached a success rate of 93,4%. The training MSE was 0.00319 and the test MSE 0.12876. The similarity in percentages between this experiment and the one with 81 neurons clearly shows the importance of certain metrics (Figure 3).

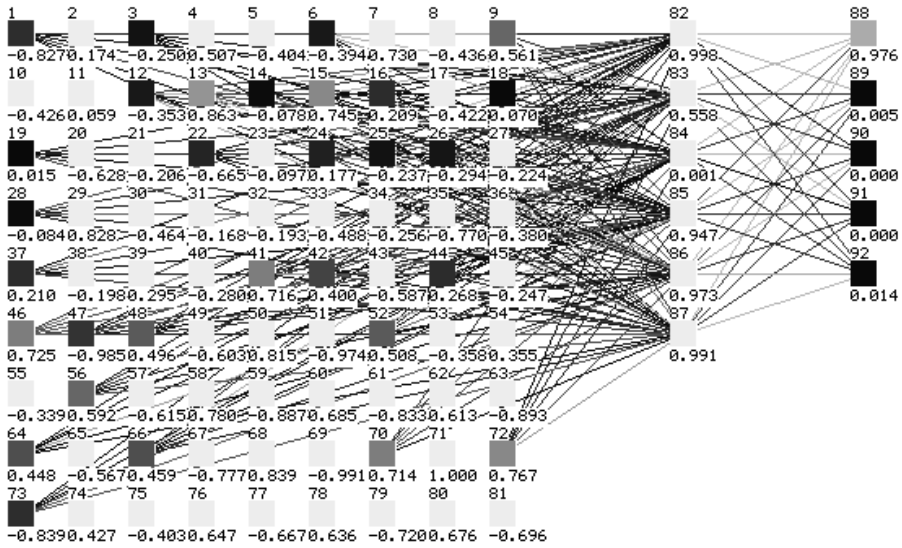

Figure 3. Architecture of the trained network with an input of 30 neurons. The image shows all the neurons (81), but only the 30 selected neurons have connections with the hidden layer (darker shade).

Architecture 15-6-5:

The selection took place in the same way as in the previous experiment, but the input layer counts 15 neurons. In this case, the results were less positive: we observed 15 errors (more than twice the previous amount), which amounts to a success rate of 85,9%. The training MSE was 0.01131 and the test MSE 0.26105.

Several representative metrics were excluded from the selection, and the 15 selected metrics were not enough to distinguish among the authors.

4. Conclusions

This paper describes a generic framework for the development of artificial art critics, including an architecture and a validation methodology.

The proposed architecture separates generic from domain specific components, allowing an easy adaptation to different domains. It also proposes a multilevel validation methodology that allows a structured testing of artificial art critics and enables the comparison of different approaches.

We have built a system that uses the proposed architecture through the combination of a static model, that analyzes the musical pieces, and an adaptive model, based on artificial neuron networks that predict the author through the output of the first model. This output corresponds to a set of 81 features that includes harmonic consonance and distance of note repetitions. We use fractal metrics to recursively applying simple metrics at decreasing resolution levels.

A high success rate shows that the system has proven its efficiency in the performance of the proposed task, which is the identification of the author of a given piece of music. The task was relatively complex, because in several cases the authors belonged to similar schools and periods.

The ANN has shown its capacity to detect the relevant metrics:

- Among all the neuron selections, the ANN always maintained more than 85% of the relevant metrics. This means that if two tests used 30 relevant metrics, approximately 25 metrics were the same in all the cases.

- We obtained similar results in the tests carried out with only 30 metrics. This shows that it is not necessary to use 81 metrics; however, the less positive results of the experiments based on 15 metrics indicate that there is a minimum of relevant metrics.

This mechanism makes it possible to incorporate metrics obtained by several research groups: the adaptive system (such as the ANN-based system used in this article) discerns the metrics that are relevant to the identification of the author.

We are carrying out more experiments to verify the ability of this system in the identification of visual arts; and we are about to test this system by means of the second level of the methodology, the subjective appraisal.

References

[1] A. Pazos, A. Santos, B. Arcay, J. Dorado, J: Romero, and J. Rodríguez. An Application Framework for Building Evolutionary Computer Systems in Music. Leonardo, 36(1), 2003

[2] S. Baluja, D. Pomerleau, and T. Jochem. Towards Automated Artificial Evolution for Computer-Generated Images. In Connection Science 6, No. 2, pp. 325–354. 1994.

[3] David Cope. Experiments in Musical Intelligence. Madison, WI: A-R Editions, 1996.

[4] Rudolf Arnheim. Entropy and Art. University of California Press, 1971.

[5] P. Machado and A. Cardoso. All the truth about NEvAr. Applied Intelligence, Special issue on Creative Systems, Bentley, P. Corne, D. (eds), Vol. 16, Nr. 2, pp. 101–119, Kluwer Academic Publishers, 2002.

[6] L. Spector and A. Alpern, Criticism, Culture and the Automatic Generation of Artworks. In Proceedings Twelfth National Conference on Artificial Intelligence (AAAI-94), August 1-4, pp. 3–8. AAAI Press. 1994.

[7] B. Manaris, D. Vaughan, C. Wagner, J. Romero and R. Davis. Evolutionary Music and the Zipf-Mandelbrot Law: Developing Fitness Functions for Pleasant Music. In Lecture Notes in Computer Science, Applications of Evolutionary Computing – EvoWorkshops 2003, LNCS 2611, pp. 522-534, Springer-Verlag, 2003