Power to the Critics – A Framework for the Development of Artificial Art Critics

Penousal Machado[1], Juan Romero[2], Bill Manaris[3], Antonino Santos[2], Amílcar Cardoso[4]
[1]Instituto Superior de Engenharia de Coimbra, Qta. da Nora, 3030 Coimbra, Portugal
[2]Creative Computer Line – RNASA Lab, Fac. of Computer Science, University of La Coruña, Spain
[3]Computer Science Department, College of Charleston, Charleston, SC 29424, USA
[4]CISUC – Centre for Informatics and Systems, University of Coimbra, 3030 Coimbra, Portugal
machado@dei.uc.pt, jj@udc.es, manaris@cs.cofc.edu, nino@udc.es, amilcar@dei.uc.pt

Abstract

In this paper we propose a framework for the development of artificial art critics, consisting of an architecture and a validation methodology. The architecture includes two modules: a feature extractor, which performs a pre-processing of the artwork, extracting several measurements and characteristics; and an evaluator, which, based on the output of the feature extractor, classifies the artwork according to a certain criteria. The validation methodology comprises several stages, ranging from author and style discrimination to the integration of the artificial art critic in a dynamic multi-agent environment, which includes human agents. Finally, we present experimental results obtained by the implementation of the proposed framework.

1 Introduction

It is common to associate the ability to generate works of art with creativity. As such, the development of computational systems that create artworks may provide significant insights and contributions to the study of creativity.

The artistic process depends highly 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, in turn, depend on the ability of seeing/listening. As Boden states:

> "someone that has a new idea must be able to evaluate it by itself" [Boden, 1990]

We think that modeling this capacity of the artist is an important, if not necessary, step in the creation of a "real" artificial artist. After all, an artist is also, and foremost, a viewer/listener.

This contrasts with the vast majority of the computational systems for artwork generation that have been developed during the past few years[1]. Typically, the role of the viewer/listener is completely neglected; such systems have neither the ability to perceive the artworks produced by them (or by other artists), nor are they able to perform aesthetic judgments. As such, these systems tend to be completely blind/deaf to the outside world.

In this paper we present a general framework for the development of Artificial Art Critics (AACs), i.e. systems that are capable to see/listen to an artwork and perform some sort of evaluation of the perceived piece. This framework, based on an analysis of existing AACs, consists of an architecture and a validation methodology.

The proposed architecture encompasses a *feature extractor* and an *evaluator* module. The *feature extractor* 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 for the *evaluator*, which assesses the artwork according to a specific criterion or aesthetics.

One of the main difficulties in the development of Computational Artists, and more specifically AACs, is their validation. To help address this problem we propose a multi-stage validation methodology. The first stage of this methodology allows the objective, and meaningful, assessment of the AACs, providing a solid basis for their development. The later stages incorporate more dynamic criteria, and include testing the AACs in a hybrid society of humans and artificial agents.

We tested our ideas by developing an AAC in the musical domain, and conducted a set of experiments, which, although preliminary, give promising results.

2 Related Work

In this section we make a brief analysis of AACs and art generation systems that incorporate art evaluation capabilities. We focus on examples from the visual arts and musical domains. We base our analysis on the approach

[1] For a survey on computational approaches to music composition a survey see, e.g., [Papadopoulos and Wiggins, 1999].

used in the creation of the AAC. The most common techniques employed are: Artificial Neural Networks (ANNs), Rule Based, and Evolutionary Computation (EC).

There are several attempts in the field of music to employ ANNs to build AACs. In the work of Burton and Vladimirova [Burton, 1996; Burton and Vladimirova, 1997] an ARTMAP ANN is used to classify rhythmic patterns. The ANN is trained with examples of rhythms from drum machines. The ANN classifies the training set in clusters or "styles". The user can add or delete clusters dynamically. The output of this critic reflects the distance of the prototype vector with respect to the most similar cluster. The results of this work are very interesting. The system was used in conjunction with a genetic algorithm that creates new pieces. The generated pieces are similar to the ones used for training the ANN, which shows that the learning task was successful. Unfortunately, this is also one of the shortcomings of the system: the generated pieces are too close to the ones of the training set, and hence of little novelty.

Brad Johanson and Riccardo Poli [1998] use two different critics to supply the fitness values in a Genetic Programming (GP) based music composition system. Both critics employ a feed-forward ANN trained through back-propagation. The training samples are generated by means of an interactive GP composition system relying on human users for fitness assignment. The first critic assigns a score to each piece of music; the second one determines the "winner" in a tournament between two songs. Although performing different tasks, the networks share the same weights. The results are interesting, showing an average error of +/-5 points in a 0-100 scale. Nevertheless, the results achieved by using the AAC to guide the evolution of the composition system are worse than the ones obtained with the interactive GP system.

Another interesting system that combines an ANN critic with a GP composer is presented in [Spector and Alpern, 1994]. In this system, a feed-forward ANN, with one hidden layer, was used to determine "reasonable continuations to reasonable fragments of jazz melodies" [Spector and Alpern, 1994]. The ANN was trained with combinations of Charlie Parker's and random fragments, assigning positive responses to good combinations and negative ones to random or reverse combinations. The results of the ANN are very promising; for instance, it assigned positive responses to pieces of Jimmy Hendrix. However, the behavior of the critic and composer system is altogether a bit "unsatisfactory". The authors were able to improve the results by combining explicit symbolic rules in conjunction with the ANN. Nevertheless, as the authors state, the long-term solution is to improve the ANN critics.

Al Biles also developed an ANN critic integrated in an evolutionary computation musical composer. The ANN has a cascade-correlation (Cascor) architecture. The training and test sets are constructed from the use of an interactive evolutionary composition system. Biles tried different representations of the musical theme as inputs of the ANN, including: statistical parameters (the number of unique new note events, the size of the maximum interval, etc.); an interval histogram; the note and interval onset structures. The results of the experiments are not completely satisfactory. As Biles states:

> "when the network was able to learn the training set, it failed on the testing set" [Biles et al, 1996].

Baluja et al. [1994] presented an AAC in the domain of visual arts. The proposed system encompasses an ANN, which performs the role of the critic, and a GP system that generates images. An interactive version of the GP system was used to construct the test and training set. Baluja et al. tested five different ANN architectures. The experimental results are "somewhat disappointing" and "very difficult to quantify" [Baluja et al., 1994]. Although the ANNs are able to demonstrate some degree of learning, they fail when used to guide evolution. According to the authors, the evolutionary process, when guided by ANNs, became "very limited and largely uninteresting". Moreover, apparently the evolutionary process is able to find loopholes in the evaluation done by the ANNs and to exploit these in order to artificially enhance fitness. It is also interesting to notice that the smallest of the tested network architectures was the one that yielded the best results when applied to the test set. This may indicate that the amount of images used to train the larger networks was too small, allowing them to do memorization, and thus preventing generalization. Overall, this study shows the difficulties involved in using ANNs to create a critic in the visual arts domain.

Saunders and Gero [2001] resort to a self-organizing Kohonen network to assess the novelty of a given image. As before, a GP system is used to generate images, and each image is classified into a category represented by one of the network neurons. As evolution progresses, the critic learns a map of the typical artworks being generated by the GP system. The novelty value of a new image is assessed by calculating the categorization error of the image (i.e. if the image does not fit well in any category it has a high novelty value). This approach is particularly interesting since it takes into account the novelty aspect of the images, thus preventing stagnation in the evolutionary process. The downside is that this is the only aspect of the image being contemplated.

Wiggins and Papadopoulos [Wiggins and Papadopoulos, 1999; Papadopoulos and Wiggins, 1998] incorporate a rule-based musical critic in an evolutionary computation composer that generates jazz "solos". The rules of the critic are based on bibliography about improvisation, informal statistical analysis of jazz solos, and intuitive author-derived heuristics. The critic results are not analyzed in detail. Nevertheless, the authors point out that the results of the evolutionary system guided by this critic are interesting but still far from ideal.

Todd and Werner [1998] present a co-evolutionary approach. In this work a population of creators and a population of critics evolve together to create new musical

forms. This work is very interesting from the point of view of the study of dynamics in a society. However, the engendered aesthetics is independent from the human aesthetics. As such, it is difficult to evaluate the results.

Greenfield [2002] proposes a co-evolutionary approach to evolve images. The idea is to evolve a population of hosts (images) and a population of parasites (filters). The images are generated by a GP system. The filters are standard digital convolution filters. The filters are applied to small portions of the images, and the outcome is compared with the original image. The fitness of the image is proportional to the dissimilarity between original and convoluted images. The fitness of the filters depends on how well they manage to "pass unnoticed". This causes an arms race between hosts and parasites. Similarly to the previous co-evolutionary approach, the experimental results are hard to evaluate. The system seems to yield images of moderate complexity that have the potential to be interesting. However, these images tend to lack structure and to have large noisy regions.

3 Framework for Development of AACs

Taking into account the state of the art in this area, we propose a framework for the development of an AAC. The main goal is to provide a solid basis for the development and validation of AACs, allowing the integration of contemporary critics, and promoting collaboration in the creation of AACs. The design of this framework is based on a set of characteristics that we consider desirable:

- Adaptability – The AACs should evolve and adapt over time. We are primarily interested in this type of AAC, because they mimic better the behavior of human critics; moreover, they can be used in different artistic styles, aesthetics and tasks.
- Sociability – Ideally, the AACs should be able to adjust their behavior according to the demands of the society in which they are integrated. That is, the AACs must be able to perform in a hybrid environment – an environment that incorporates humans and artificial systems. So, the AAC must be validated by the society of artificial and human "agents" in the same way that human critics are validated in purely human societies [Pazos et al, 2003].
- Generality – The framework should allow the development of AACs for different domains; the domain-specific tasks should be carried out by specialized modules, enabling an "easy" adaptation of the AAC to new domains.
- Independence of Representation – The AACs should be able to perceive the artworks. Thus, the AACs should form their assessment from the artwork itself, without access to additional information. The AAC should only have access to the piece of art. It may build its own internal representation of the artwork, but it cannot access any sort of high level representation originally used in its construction.

3.1 Architecture

Taking into account the above characteristics, it is clear that the architecture must allow the development of generic and adaptive AACs. This quest for generality clashes with the need for handling the particularities of each domain. There are, for instance, significant differences between the way music and visual art are experienced. The most obvious difference is that music follows a predetermined temporal sequence, while in visual art the viewer has direct access to all the regions of the painting[2]. Therefore, specific modules should carry out tasks that are particular to a domain, promoting the generality of the remaining modules.

The amount of information contained in some artworks is huge. In visual art, for instance, even a relatively small picture can fill a lot of memory. Taking into account the current state-of-the-art in adaptive systems (e.g. neural networks, genetic algorithms) it is clear that these techniques cannot currently handle such vast amounts of information.

To tackle this problem some researchers resort to reducing the size of the artworks fed into the adaptive system (e.g. [Baluja et al., 1994] use 48*48 pixel images). This approach, however, poses significant problems, since a considerable amount of detail may be lost. Moreover, the experimental results are, typically, disappointing.

We think that it is more adequate to preprocess the artworks, in order to extract relevant features. These features would serve as input for the adaptive part of the system, thus significantly reducing the amount of information.

With this set of ideas in mind we developed a generic architecture for an AAC.

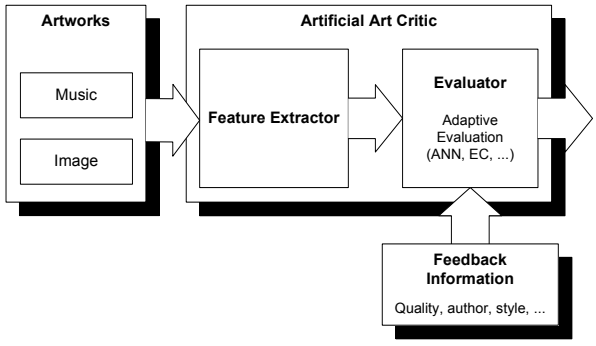

Figure 1. Outline of the proposed model.

[2] As time progresses the viewer may focus on particular regions of the image, but the sequence is not predetermined. Although an experienced artist can guide the eye of the viewer, the process of perception seems less constrained in visual art than in music.

The proposed architecture encompasses two modules: the *feature extractor* and the *evaluator*. The role of the *feature extractor* is to perform an analysis of the artwork providing a set of relevant features to the *evaluator*. The role of the *evaluator* is to make an assessment of the artwork based on the supplied features. Figure 1 presents a rough outline of the proposed architecture.

The feature extractor module typically comprises two tasks: *perception* and *analysis*.

In the *perception* task the system converts the artwork into a "percept", i.e., some sort of internal representation of the perceived reality. The outcome of the perception task is then analyzed, which yields a set of relevant features or measurements. In some cases the division between perception and analysis is more conceptual than real. The main idea is that the perception acquires information about domain-specific parameters, which are then analyzed.

We do not impose any kind of constraints on the type of internal representation, nor on the range of techniques used on the *feature extractor*. Therefore, this can include statistical, rule-based, algorithmic, symbolic, sub-symbolic techniques, etc.

The *evaluator* module is an adaptive system that takes as input the characterization of the artwork done by the *feature extractor*, and outputs an assessment of the artwork.

In this framework the *evaluator* module is not specifically designed to a particular task. It should have enough flexibility to adapt to different tasks according to the feedback information provided to it (see figure 1). This information indicates the desired answer or, depending on the task at hand, an evaluation of the performance of the AAC. The adaptive *evaluator* should adjust its behavior, according to the received feedback, in order in order to maximize its performance.

The adaptive *evaluator* module can give information about the features that are relevant in the assessment of an artwork. As an example, the weights of an ANN can show what features are taken into account, indicating what the most relevant features for a particular task. It is also possible to find the minimum set of features necessary to achieve some task by testing the *evaluator* with different sets of features.

The proposed architecture allows a certain degree of independence between the search for new relevant features and evaluation. It is possible to include in a system a set of *feature extractors* from different authors, in order to test which one characterizes the piece better. The same can be applied to the *evaluators*. The data from the *feature extractor* and the feedback information can be applied to different *evaluators* in order to compare them.

Next we present a validation methodology that was designed to allow a structured testing of the developed AAC.

3.2 Validation Methodology

The validation of an AAC poses difficult problems. These are mostly related to the subjectivity involved in the evaluation of artworks. Additionally, a large training set is required to train the *evaluator* module. This implies having hundreds, or even thousands, of human evaluated artworks, which, needless to say, can be a problem in itself.

In order to begin addressing these difficulties we propose a multi-stage validation methodology. In each level the AAC is presented with a different task. We start with tasks in which the correctness of the AACs output can be objectively determined, and which do not require a training set of human evaluated artworks. Then we move on to tasks of higher subjectivity and complexity. In the first levels the response of the AAC is supposed to be static. In the latest level, however, the AAC is required to adapt to the environment and to change its evaluation over time according to the surrounding context. Currently, we consider three levels of validation: Identification, Static Evaluation, and Dynamic Evaluation.

The identification level deals with assessing the ability of the AAC to recognize the style or author of a given artwork.

In 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 type of validation is relatively easy to perform, the compilation of training instances is straightforward, and the test is totally objective. The main difficulty involved in this level of testing is the construction of representative training and test sets.

Although limited in scope, this validation step is useful in determining the capabilities of the *feature extractor* module. A failure in this test may indicate that the set of extracted features is not enough to discriminate between authors, thus, preventing us to move to a more complex task, which is bound to fail due to the lack of meaningful information. Moreover, an analysis of the features used by the *evaluator* to determine the right author can help determine the relative importance of each of the extracted features. In fact, one can perform specific tests to determine the predictive power of each measurement or set of measurements.

The task of *Style Identification* is similar to the previous one. The difference is that in this case the AAC must identify the style of an artwork. The training and testing can be done pretty much in the same way as in the *Author Identification* step. This type of validation allows the testing of AACs that may be used in a wide variety of tasks, such as image and musical retrieval, allowing style-based searches.

The overall and relative difficulty of these tasks depends on the chosen artists and styles. Trying to discriminate between artists of the same school can be more difficult than distinguishing radically different styles. However, discriminating between artists that have char-

acteristic signatures (in the sense used by Cope [Cope, 1996]) is easier than discriminating between closely related styles. In 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 second level of validation is the *Static Evaluation*. The task of the AAC is to determine the aesthetic value of a series of artworks previously evaluated by humans. One of the major difficulties in performing this test is the construction of a representative database of consistently evaluated artworks.

It is important to notice that the training of the AAC requires not only positive examples but also negative ones. Thus, a vast amount of bad pieces are needed. Ironically, it is quite difficult to get a representative set of the "wrong things to do".

Moreover, one also needs a representative sample of items that do not even meet the necessary requirements to be considered a piece, e.g. images in which the pixels are totally uncorrelated, and, as such, are nothing more than noise. The use of complexity appraisers in the *feature extractor* module may prove useful to rule out this type of items. The relation between complexity and aesthetic value has been pointed out by several authors (see, e.g., [Arnheim, 1971]); and complexity appraisers have successfully been used as a way to filter images that do not meet the necessary pre-requirements to be considered artworks [Machado, 2002].

To create the training set, one can resort to a generative art tool. This would yield a relatively high number of pieces in a reasonable amount of time. However, the consistency of the evaluation depends vastly on the discipline of the user. Additionally, the set will only be representative of the pieces 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; that is to 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. For instance in [Teller and Veloso, 1995] the authors trained a face recognition system, which had surprisingly good performance. However, a careful analysis of the experimental results showed that the system was not recognizing the faces of the people in the images, it was recognizing the offices in which the pictures where taken.

To detect this type of problem, we suggest 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 typically good at exploiting holes in the fitness evaluation (see, e.g., [Spector and Alpern, 1994]). 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 on 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. Thus, its behaviour must be socially adequate. To perform this validation, a model of society named "Hybrid Society" (HS) is proposed. HS is a paradigm similar to Artificial Life, but with human "agents" at the same level of artificial ones. HS explores the creation of egalitarian societies populated by humans and artificial beings in artistic (or other social) domains[3]; 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.

The problem of this validation level is the need to incorporate humans in the experimentation. So, 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.

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

The validation methodology presented here tries to find a compromise between automated and human-like validation. We are fully aware of the difficulty of the proposed tasks. It is important to notice, however, that

[3] For more information about the Hybrid Society Project see http//www.hybridsociety.net

for certain tasks you only need to take into account some of the validation levels.

4 Experimental Results

Using the presented framework we developed an AAC in the musical domain. We conducted a set of experiments, which correspond to the first validation level. The task presented to the AAC was to discriminate between musical themes from two authors, Beethoven and Bach. We used 108 scores from Bach and 32 from Beethoven.

Following the proposed architecture, the system has two modules, the feature extractor and the adaptive evaluator, which are described in the following sections.

4.1 Feature Extractor

The static feature extractor has been described in [Manaris et al., 2002]. It utilizes a collection of metrics, based on Zipf's law, to extract a series of features from music pieces encoded in MIDI.

Zipf's law states that natural phenomena generated within societies of interacting, self-adapting organisms follow *the principle of least effort* [Zipf, 1949]. This principle dictates that over time, which may span several generations, such organisms will attempt to minimize the probable average rate of their work expenditure. This produces a fascinating by-product: The phenomena generated by the interaction of such organisms (e.g., language) exhibit a predictable structural equilibrium. This equilibrium is formulated in terms of the frequency of occurrence of events within such phenomena (e.g., words within a book), as follows:

$$P_n \sim 1/n^a \quad (1)$$

where P_n is the frequency of occurrence of the n^{th} ranked item, and a is close to 1. This is known as a Zipf distribution. A more general form of (1), known as a Zipf-Mandelbrot distribution, is:

$$P_n \sim (1+b)/n^{(1+c)} \quad (2)$$

where b and c are arbitrary real constants.

Zipf distributions have been discovered in a wide range of man-made and naturally occurring phenomena including city sizes, salaries, word frequencies, earthquake magnitudes, thickness of sediment depositions, extinctions of species, traffic jams, and visits of websites [Li, 2003]. They have also been discovered in natural language (both written and spoken) as well as music.

The following studies motivated the adoption of Zipf-based metrics in the construction of the feature extractor. Zipf [1949, pp. 336–337] shows that four popular music pieces exhibit this equilibrium, in terms of melodic intervals, and of distance between repetitions of notes. Zipf's distributions are also found in the pitch and loudness fluctuations of radio signals recorded over a 24-hour period from classical, jazz, blues, and rock stations [Voss and Clarke, 1975]. Finally, a study of 220 pieces of various music styles (baroque, classical, romantic, twelve-tone, jazz, rock, DNA strings, and aleatory music) discovers many additional Zipf distributions [Manaris et al., 2003]. This study also shows that, for certain characteristic styles such as twelve-tone and aleatory music, style identification is possible through simple statistical means, such as ANOVA analysis. Another interesting result is that Zipf distributions across certain music attributes appear to be a necessary, but not sufficient condition for pleasant music. This implies that Zipf-based metrics could be used for evaluating the promise of early drafts of artifacts, e.g., a music theme to be subsequently incorporated into a larger work.

The feature extractor used in this experiment employs Zipf-based metrics on several music-theoretic and information-theoretic attributes. Music-theoretic attributes include pitch, pitch and duration, melodic intervals, and harmonic intervals. Information-theoretic attributes include melodic and harmonic bigrams, melodic trigrams, higher-order melodic intervals and fractal aspects of musical balance. Each metric 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 mean square error, R^2, of the trendline; this ranges from 0 to 1, with 1 denoting a perfect fit.

In this experiment, the feature extractor produced 15 metrics – a total of 30 values – from each music piece.

For example, Figure 3 shows the distribution of melodic intervals for Chopin's "Revolutionary Etude," Opus 10 No. 12 in C minor. The slope is near-Zipfian (-1.1829) with an al-most-perfect fit (0.9156).

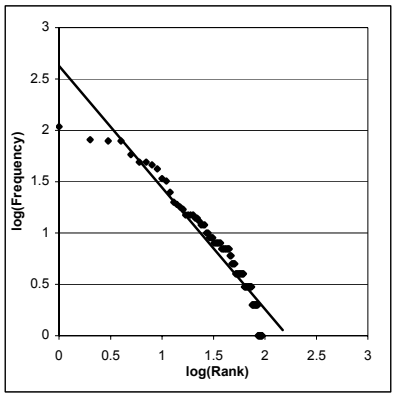

Figure. 3. Distribution of melodic intervals for Chopin's "Revolutionary Etude", Opus 10 No. 12 in C minor. Slope of the trendline is -1.1829, R2 is 0.9156.

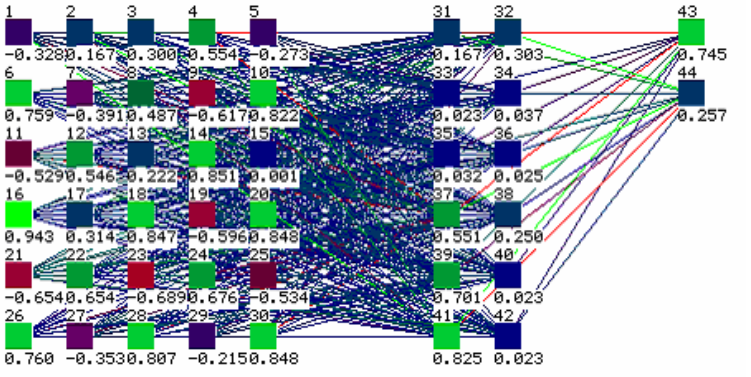

Figure 2. Architecture of the trained ANN.

4.2 Adaptive Evaluator

In this section we describe the adaptive evaluator used in this experiment. The adaptive evaluator consists of a feed-forward ANN with one hidden layer. After some experimentation with different ANN architectures, we chose one with 30 elements in the input layer, 12 in the hidden layer, and 2 in the output layer. Each unit in the input layer corresponds to each of the values generated by the metrics. These values where normalized to fit the [-1, 1] interval. An output of (1,0) indicates that the author of the score is Beethoven, while (0,1) indicates a Bach score.

The training set contained 66% of the scores of each composer, the scores were randomly selected; the testing set contained the remaining ones.

We used SNNS[4] to build, train and test the ANN. As learning function we chose back-propagation. The trained ANN is presented in figure 2.

The training of the ANN took 20000 cycles. Figure 4 shows the evolution of the error on the training and test sets as the learning progressed. The learning rate was set to 0.1 and momentum to 0. In this experiment the ANN successfully identified the authors of all the scores in the test (and training) set. There is, however, a seemingly atypical score: in the experiments where it wasn't included in the training set, the ANN failed to identify its author correctly.

Upon completion of the experiment, we tried to identify the features that are most relevant for the recognition of the author. To achieve this, we resorted to an analysis of the ANN connection weights. In figure 5 we present their values after training. Some of the processing elements of the input level have weights significantly higher than others. This indicates that the associated features are more relevant. These features are, by order of relevance: R^2 of pitch relative to octave; R^2 of pitch; slope of pitch; slope of harmonic and melodic intervals; and R^2 of the sixth-order melodic intervals.

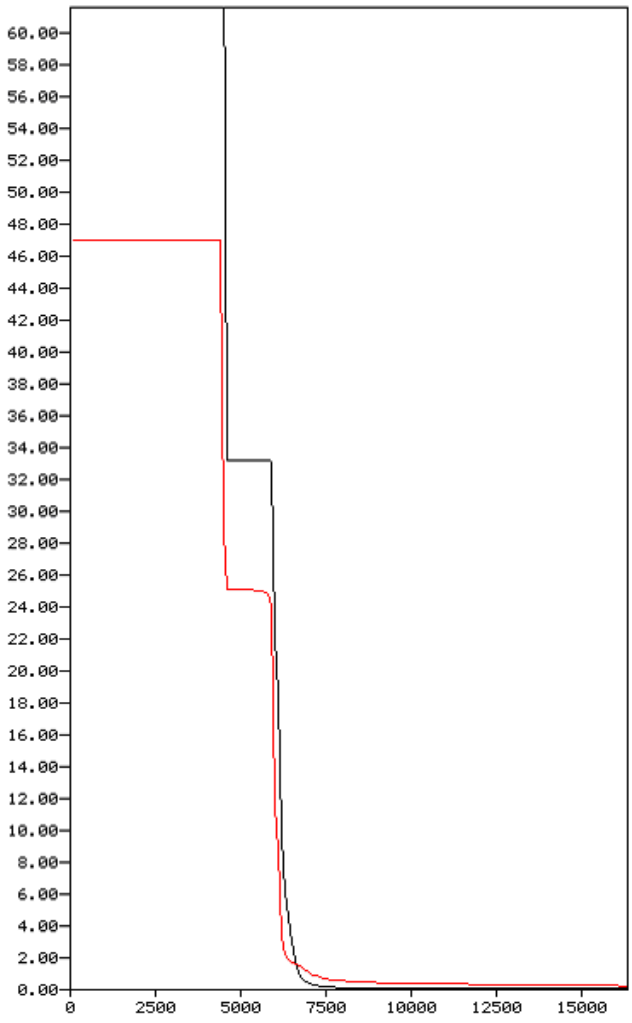

Figure 4. Error of the ANN on the training and test sets. The grey line (that begins in 47) shows the error on the test set; the black line corresponds to the error on the training set.

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

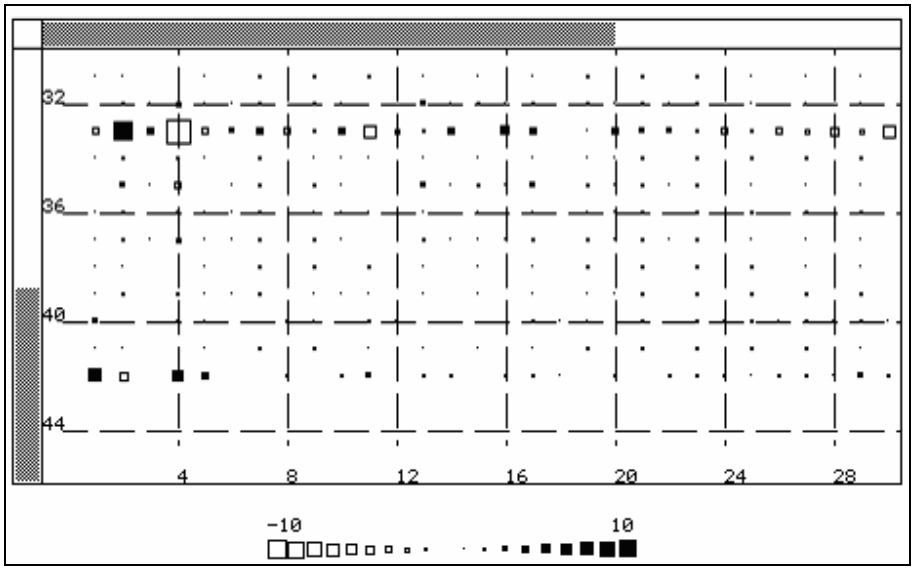

Figure 5. Connection weights between the input processing elements (X axis) and the elements of the hidden layer (Y axis) of the ANN.

We performed several repetitions of the experiment. All the repetitions indicated the same set of relevant features. Additionally, we performed several experiments in which only the most relevant features were used. This introduced only a small degradation of performance, suggesting that this reduced set of features is indeed the most relevant for the task of discriminating between these two authors. Clearly, other tasks may correspond to a different set of relevant features.

5 Conclusions and Further Work

We propose a generic framework for the development of artificial art critics, based on an analysis of the current state of the art in the area, and on the experience acquired in the development of previous systems. This framework includes an architecture and a validation methodology.

In order to allow an easy adaptation to different domains, the proposed architecture separates generic from domain specific components. Furthermore, it also establishes a boundary between static and adaptive modules. The validation of an artificial art critic is a complex task. We propose a multilevel validation methodology that allows a structured testing of artificial art critics and enables the comparison of different approaches.

Following the proposed framework, we implemented an artificial art critic, and tested its performance in the task of author identification. This yielded promising results.

Currently, we are developing an artificial art critic in the visual arts domain. We are also conducting a new array of experiments in the music domain with an extended ANN architecture. This architecture accommodates a wide variety of simple and fractal metrics – a total of 80 features. Simple metrics, in addition to the ones mentioned above, include harmonic consonance and distance of note repetitions. Fractal metrics apply simple metrics recursively at decreasing levels of resolution within a music piece, and report the corresponding fractal dimension. Future steps involve testing our approach at the other levels of the proposed validation methodology.

An interesting possibility is to explore whether Zipf's principle of least effort could be used, at a higher-level, to evaluate the efficacy and "naturalness" of an arbitrary egalitarian society of AACs by examining various aspects of societal interaction among agents. For example, we could examine the distribution of worth among individual art critics in the society – how "respected" the "opinion" of an individual art critic may be among its peers; this is analogous to the notion of salary in human societies, which has been shown to obey Zipf's law. Several other analogies may be drawn from natural phenomena that have been shown to follow Zipf's law [Li, 2003].

The AAC framework described herein is not constrained to strictly artistic domains. It can be used in any domain that involves (a) the creation of a hypothesis (design, solution, etc.), and (b) the iterative refinement of that hypothesis based on aesthetics, constraints, and other quantifiable attributes. Such domains include software development, mathematics, engineering, and architecture. For instance, Zipf metrics have already been used to evaluate software, architectural design and other complex systems [Shooman and Laemmel, 1977; Salingaros and West, 1999]. Additionally, the application to other areas such as content-based image and music retrieval also seems viable.

Research in the area of artificial art critics and artists is still in an embryonic stage. The proposed framework is intended to provide a common foundation for the devel-

opment and validation of artificial art critics, and to promote collaboration among researchers in this area.

The construction of artificial artist critics is an important step in the design of a true artificial artist, and potentially in the better understanding of the artistic and creative process.

Acknowledgments

We take this opportunity to express our gratitude to Robert Davis for various statistical analyses; Charles McCormick, Tarsem Purewal, Dallas Vaughan and Christopher Wagner for contributing to the development of Zipf-based metrics; and, last but not least, Marisa Ares for contributing to the implementation of the evaluator module and for providing comments on an earlier draft of this paper.

This work was partially supported by the Portuguese Ministry of Education, program PRODEP III, Action 5.3, by the Calouste Gulbenkian Foundation, and by an internal research grant from the College of Charleston.

References

[Arnheim, 1954] Rudolf Arnheim. *Art and Visual Perception: A Psychology of the Creative Eye*. University of California Press, 1954.

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

[Baluja et al, 1994] 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.

[Biles et al, 1996] J. A. Biles, P. G. Anderson and L. W. Loggi. Neural Network Fitness Functions for an IGA. In *Proceedings of the International ICSC Symposium on Intelligent Industrial Automation (ISA'96) and Soft Computing (SOCO'96)*, 26–28 March 1996 (Reading, U.K.: ICSC Academic Press, 1996) pp. B39–B44. 1996.

[Boden, 1990] M. A. Boden. *The Creative Mind: Myths and Mechanisms*. London, Cardinal. 1990.

[Burton, 1996] A. R. Burton. *A Hybrid Neuro-Genetic Pattern Evolution System Applied to Musical Composition*. PhD thesis, University of Surrey. 1996.

[Burton and Vladimirova, 1997] A. R. Burton and T. Vladimirova. Genetic Algorithms Utilizing Neural Network Fitness Evaluation for Musical Composition. In G.D. Smith, N.C. Steele and R.F. Albrecht (Eds.), *International Conference on Artificial Neural Networks and Genetic Algorithms,* Norwich, UK, pp. 220–224, Springer, 1997.

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

[Greenfield, 2002] Gary R. Greenfield. On the Co-Evolution of Evolving Expressions. *International Journal of Computational Intelligence and Applications*, Vol. 2, No. 1 (2002) pp. 17–31, Imperial College Press, 2002.

[Johanson and Poli, 1998] Brad Johanson and Riccardo Poli. GP-Music: An Interactive Genetic Programming System for Music Generation with Automated Fitness Raters. *Genetic Programming 1998: Proceedings of the Third Annual Conference*, 1998

[Li, 2003] Wentian Li. Zipf's Law, http://linkage.rockefeller.edu/wli/zipf/. Acessed May, 2003.

[Machado and Cardoso, 2000] 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.

[Manaris et al., 2002] B. Manaris, T. Purewal and C. McCormick. Progress Towards Recognizing and Classifying Beautiful Music with Computers-MIDI-Encoded Music and the Zipf-Mandelbrot Law. In *Proceedings of EEE SoutheastCon 2002*, Columbia, SC, pp. 52–57, 2002.

[Manaris et al., 2003] 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.

[Papadopoulos and Wiggins, 1998] G. Papadopoulos and and G. A. Wiggins. A Genetic Algorithm for the Generation of Jazz Melodies. In *Proceedings of STeP'98: 8th Finnish Conference on Artificial Intelligence.* Jyväskylä, Finland, September 7-9, 1998.

[Papadopoulos and Wiggins, 1999] G. Papadopoulos and and G. A. Wiggins. AI Methods for Algorithmic Composition: A Survey, A Critical View, and Future Prospects. *Proceedings of the AISB'99 Symposium on Musical Creativity*, 1999.

[Pazos et al, 2003] 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

[Salingaros and West, 1999] N. A. Salingaros and B. J.West. A Universal Rule for the Distribution of Sizes. *Environment and Planning,* B(26), pp. 909–923, 1999.

[Shooman and Laemmel, 1977] M. Shooman and A. Laemmel. Statistical Theory of Computer Programs. In *Proceedings of IEEE Computer Conference*, pp. 511–517, Oct. 1977.

[Saunders and Gero, 2001] R. Saunders and J. Gero, Artificial Creativity. A Synthetic Approach to the Study of Creative Behaviour. In J. S. Gero (ed.), *Proceedings of the Fifth Conference on Computational and Cognitive Models of Creative Design*, Key Centre of Design Computing and Cognition. 2001.

[Spector and Alpern, 1994] 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.

[Teller and Velosos, 1995] A. Teller and M. Veloso. Algorithm evolution for face recognition: What makes a

picture difficult. *In Proceedings of the International Conference on Evolutionary Computation*, IEEE Press, 1995.
[Todd and Werner, 1998] P. M. Todd and G. M. Werner. Frankensteinian Methods for Evolutionary Music Composition. In N. Griffith and P.M. Todd, Eds., *Musical Networks: Parallel Distributed Perception and Performance*, Cambridge, MA, MIT Press, 1998.
[Voss and Clarke, 1975] Voss, R.F., and Clarke, J.: 1/f Noise in Music and Speech. *Nature* 258, pp. 317–318, 1975.
[Wiggins and Papadopoulos, 1999] G. A. Wiggins, G. Papadopoulos, S. Phon-Amnuaisuk and A. Tuson. Evolutionary Methods for Musical Composition. In *International Journal of Computing Anticipatory Systems*, 1(1), 1999.
[Zipf, 1949] G. K. Zipf. *Human Behavior and the Principle of Least Effort*. New York: Hafner Publishing Company, 1949.