Chapter 11
Computing Aesthetics with Image Judgement Systems

Juan Romero, Penousal Machado, Adrian Carballal, and João Correia

Abstract The ability of human or artificial agents to evaluate their works, as well as the works of others, is an important aspect of creative behaviour, possibly even a requirement. In artistic fields such as visual arts and music, this evaluation capacity relies, at least partially, on aesthetic judgement. This chapter analyses issues regarding the development of computational systems that perform aesthetic judgements focusing on their validation. We present several alternatives, as follows: the use of psychological tests related to aesthetic judgement; the testing of these systems in style recognition tasks; and the assessment of the system's ability to predict the users' valuations or the popularity of a given work. An adaptive system is presented and its performance assessed using the above-mentioned validation methodologies.

11.1 Introduction

Creativity is frequently associated with the capacity to create artworks. Therefore, the design of computing systems which have the skills to create artworks can provide interesting insights into a general understanding of creativity. Spector and Alpern (1994) define a "Constructed Artist" as an entity that is "... supposed to be capable of creating aesthetically meritorious artworks on their own, with minimal

J. Romero (✉) · A. Carballal
Faculty of Computer Science, University of A Coruña, Campus de Elviña, CP 15071, A Coruña, Spain
e-mail: jj@udc.es

A. Carballal
e-mail: adrian.carballal@udc.es

P. Machado · J. Correia
Department of Informatics Engineering, University of Coimbra – Polo II, 3030-290 Coimbra, Portugal

P. Machado
e-mail: machado@dei.uc.pt

J. Correia
e-mail: jncor@student.dei.uc.pt

J. McCormack, M. d'Inverno (eds.), *Computers and Creativity*,
DOI 10.1007/978-3-642-31727-9_11, © Springer-Verlag Berlin Heidelberg 2012

human intervention", as opposed to other computational systems performing artistic tasks. Artistic processes often rely on the capacity to make aesthetic judgements, using artworks created by others as sources of inspiration and making criticism of their own work. As Boden (1990) puts it: "Someone that has a new idea must be able to evaluate it by itself".

A major obstacle in developing constructed artists is the difficulty of implementing aesthetic judgement mechanisms. Having a system capable of creating its own aesthetic preferences, or acquiring them from a cultural environment, would be an important step towards the development of computational creativity.

The concepts of art and aesthetics are deeply related. Nevertheless, it is important to differentiate between them. The artistic value of an artwork depends on several factors, including form, content, cultural context and novelty. We acknowledge the relevance of all these factors, yet, we focus exclusively on the aesthetic properties of the artworks, and—for the scope of this chapter—we define Aesthetics as the study of the form in itself, i.e. stripped from content, context, and all the other factors that, although relevant from an artistic standpoint, do not result exclusively from form and, consequently, cannot be analysed when considering only the form.

By assuming this point of view, we are not creating a false dichotomy between "form" and "content". We acknowledge that these factors are not independent. Form affects, and sometimes determines, the way content is perceived and conveyed, and the coherence or contrast between form and content can be explored. For instance, an artist may choose to use a composition that he finds visually pleasing and harmonious to convey content that is highly displeasing and violent, exploring the discrepancy between form and content for artistic purposes. Even when the artwork is purely abstract, one cannot rule out the possibility that a human observer perceives, even if only at a subconscious level, some type of content that evokes feelings and emotions and that, therefore, influences his reaction to the piece. In other words, it may be impossible for a human to focus exclusively on the form, which makes the discipline of aesthetics (as defined here) an unreachable goal. Although this constitutes an obvious drawback, it is also an opportunity: computers can focus exclusively on the form.

In the same way that we differentiate between Art and Aesthetics, we also differentiate between Artistic and Aesthetic Judgement. The existence of universal aesthetic preferences shared among all humans, the existence of shapes that are inherently pleasing or displeasing, the way culture and training affect aesthetics, etc. are controversial (even among the authors of this chapter). These questions, although relevant, are outside the scope of what we describe here. We consider, however, that there are properties such as symmetry, balance, rhythm, contrast, proportion, repetition, unity, predominance, variety, and continuity which are aesthetically relevant and that can be considered aesthetic principles. This does not imply that a symmetric image is inherently more pleasing than an asymmetric one. It does, however, imply that symmetry may influence the aesthetic value of an artwork. The way a given aesthetic property influences aesthetics depends on a wide variety of issues, including the relationship with other aesthetic properties, personal preferences, aesthetic trend, and so on.

We posit that the ability to recognise at least some of these aesthetic properties is common to all humans, acknowledging that the way different humans may react to different aesthetic principles, to their relationships, and value aesthetic principles may vary. Likewise, the degree of awareness to principles of aesthetical order and the inclination to use aesthetic criteria when valuing artefacts also differs.

In Machado et al. (2003) we find the following definition: Artificial Art Critics are "systems that are capable to see/listen to an artwork and perform some sort of evaluation of the perceived piece". Unfortunately, the term "art critic" can be easily misunderstood, given that it may be perceived as the equivalent of a human making an artistic critique or a written analysis of an artwork, rather than an aesthetic judgement. For this reason, we abandon this nomenclature.

Taking all of the above into consideration, for the scope of this chapter, we define an aesthetic judgement system (AJS) as a system that performs an aesthetic assessment of an image based on its aesthetic properties. For instance, a system that: measures the degree of accordance of an artwork with a given aesthetic theory; measures several aesthetic properties of an image; makes an assessment of an artwork according to the aesthetic preferences of a given user, set of users, community, etc.; identifies the aesthetic current of an artwork; assesses the aesthetic consistency of a set of works; etc.

It is important to note that the system should make its judgement based on aesthetic properties. A system that assesses the aesthetic value of an artwork by analysing its aesthetic properties can be considered an AJS. A system that performs the same task by using optical character recognition to identify the signed name of the author and determines aesthetic value by the popularity of the author cannot be considered an AJS.

An AJS may provide a quantitative judgement, e.g. a single numeric value, a vector, or a classification in one or more dimensions. An AJS may also provide a qualitative assessment or assessments. Ultimately, the adequacy of the output depends on the task at hand. For instance, to guide an evolutionary algorithm using roulette wheel selection, a quantitative judgement, or one that can be converted to quantities, is required. However, to guide the same algorithm using tournament selection, only a qualitative assessment is needed, i.e. knowing if a given individual is better suited to the task at hand than another, we do not need to quantify how much better it is.

The AJSs can be divided into two categories. The first category explores systems that rely on a theory of visual aesthetics and use an AJS to explore this theory by computing it, e.g. Rigau et al. (2008), Staudek (2002; 2003), Taylor et al. (1999), Machado and Cardoso (1998), Spehar et al. (2003), Schmidhuber (1997; 1998; 2007), see also the chapters by Galanter (Chap. 10) and Schmidhuber (Chap. 12) in this volume.

The second category presents learning systems which include some kind of adaptive capacity that potentially allows them to learn user preferences, trends, aesthetic theories, etc. Although there are different approaches, usually these systems extract information from images (e.g. a set of metrics) using a machine learning system that performs an aesthetics-based evaluation or classification. There are numerous examples of this architecture in the fields of content based image retrieval and computer vision, such as Datta et al. (2006; 2008), Ke et al. (2006), Cutzu et al. (2003). One of the advantages of this kind of systems is their potential use to perform different tasks, and to be adapted to different aesthetic preferences. Classification tasks are particularly useful for validation purposes since they tend to be objective and allow a direct comparison of the results obtained by several systems (provided that they are applied to the same datasets).

Relatively few attempts have been made in the visual arts field to integrate evaluation skills into an image generation system. Neufeld et al. (2007) presented a genetic programming engine generating non-photorealistic filters by means of a fitness function based on Ralph's bell curve distribution of colour gradient. This model was implemented by carrying out an empirical evaluation of hundreds of artworks. Their paper contains examples of some of the non-photorealistic filters created.

Kowaliw et al. (2009) compared biomorphs generated in three different ways: at random, through interactive evolution, and through evolution guided by a set of image metrics used in content based image retrieval. They compared the results of the three methods taking into account a model of creativity explained in Dorin and Korb (2009), coming to the conclusion that automatic methods gave rise to results comparable to those obtained by interactive evolution.

Baluja et al. (1994) used an artificial neural network trained with a set of images generated by user-guided evolution. Once trained, the artificial neural network was used to guide the evolutionary process by assigning fitness to individuals. Although the approach is inspiring, the authors consider the results somewhat disappointing.

Saunders (2001) used a similar approach, proposing the use of a Self Organising Map artificial neural network for the purpose of evolving images with a sufficient degree of novelty. This approach is restricted to the novelty aspects of artworks.

Svangård and Nordin (2004) made use of complexity estimates so as to model the user's preferences, implying that this scheme may be used for fitness assignment. The authors introduced some experiments in which they used sets of two randomly generated images, and compared, for each pair, the system's choices with those made by the user. Depending on the methodology used, the success rates ranged between 34 % and 75 %. Obviously, a result of 35 % is very low for a binary classification task. No example of the images considered was presented, which makes it impossible to evaluate the difficulty of the task and, as such, the appropriateness of the methodologies that obtained the highest averages. Additional information on the combination of AJSs in image generation systems can be found in Chap. 10 in this volume.

Although the integration of AJSs in image generation systems is an important goal, having autonomous, self-sufficient AJSs presents several advantages:

- It allows one to assess the performance of the AJSs independently, providing a method for comparing them. This allows a more precise assessment of the AJS abilities than possible when comparing AJSs integrated with image generation systems, since the strengths and weaknesses of the image generation systems may mask those of the AJS;

- It fosters cooperation among different working groups, allowing, for instance, the collaboration between research groups working on the development of AJS and groups that focus on the development of image generation systems;
- The same AJS may be incorporated with different systems allowing it to be used for various creativity supporting tasks.

This chapter focuses on AJS validation. The next section discusses some of the issues related to AJS validation and presents several validation methods based on psychological tests, users' evaluations, and stylistic principles. Section 11.3 describes the evolution of an AJS through time, from a heuristic based system to a learning AJS. The results obtained in the system validation by means of the approaches proposed in Sect. 11.2 are presented and analysed. Finally, we draw overall conclusions and indicate future work.

11.2 Validation Approaches for AJS

Performance comparison of two AJSs is a delicate task. The existence of a validation task to which both can be applied is a prerequisite for comparison. Unless the systems are applicable to the exact same task (which includes using the same datasets) the comparison may lead to erroneous conclusions. The validation method must be reproducible and the results should be numerically quantifiable. All components of the validation task (e.g. datasets) should be made accessible to the research community. Furthermore, it is also recommended that the datasets come from an external source (i.e. that they are not specifically made for a given AJS) and have an unbiased character. There are tasks, e.g. author identification, that despite not being directly related to the ability to make aesthetic assessments, can be useful due to their objectivity and can, potentially, complement other validation methods.

The characteristics of human aesthetic preferences—e.g. subjectivity, individuality, cultural biases, change through time, etc.—create an additional difficulty. Similarly, the interpretation of the results is also problematic and, in many circumstances, it is difficult to determine what constitutes a good result.

In this section we will explore three different ways to validate an AJS: based on psychological tests related to aesthetics, based on user evaluation, and based on stylistic classification.

11.2.1 Psychological Tests

There are several psychological tests aimed at measuring and identifying aesthetic preferences (Burt 1933) and aesthetic judgement (Savarese and Miller 1979, Furnham and Walker 2001). Some of them are employed on professional guidance, together with other psychological tests, in order to advise students about potential careers.

From the point of view of an AJS validation, they constitute a good reference, since they are relatively easy to apply and provide reproducible and quantifiable results. They also allow the comparison of the "performance" of the computer system with human evaluation, although this comparison is extremely delicate.

We will make a short analysis of two tests that are potentially useful for AJS validation, namely the *Visual Aesthetic Sensitivity Test* of Götz et al. and Maitland Graves' *Design Judgment Test*. Nadal (2007) provides further analysis of these and other psychological tests.

The Visual Aesthetic Sensitivity Test (VAST)—created by Götz (an artist) and Eysenck (Eysenck et al. 1984, Götz 1985, Eysenck 1983)—consists of a series of 50 pairs of non-representative drawings. In each pair the subject has to express an opinion as to which is the most harmonious design. Götz drew the "harmonious" designs first and then altered them by incorporating changes that he considered faults and errors according to his aesthetic views. The validity of the judgement was tested by eight expert judges (artists and critics), making preference judgements and only accepting pairs of designs on which agreement among judges was unanimous. When groups of subjects are tested, the majority judgement agrees with the keying of the items, which supports the validity of the original judgement.

There are easy, middle and difficult item levels. The difficulty level of items is established in terms of the percentage of correct responses; the more subjects give the right answer, the easier the item. Different groups of subjects, differing in age, sex, artistic training, cultural background, and ethnicity have produced very similar difficulty levels for the items. "The instructions of the test did not emphasise so much the individual's preference for one item or the other, but rather the quality of one design" (Eysenck 1983). The task is to discover which of the designs is the most harmonious and not which designs are the most pleasant. The images resemble abstract art, minimising the influence of content on preference. There was some cross-cultural comparison employing the VAST test. Iwawaki et al. (1979) compared Japanese and English children and students. Frois and Eysenck (1995) applied the test to Portuguese children and Fine Arts Students.

Graves (1946) presented "The Design Judgment Test" (DJT).[1] It was designed to determine how humans respond to several principles of aesthetic order, presented in his previous work (Graves 1951). It contains 90 slides with pairs or triads of images. In each of the slides, one particular image "is considered 'right' (and scored accordingly) on the basis of agreement with the author's theories and the agreement of art teachers on the superiority of that particular design" (Eysenck and Castle 1971). Thus, on each slide, one of the images follows the aesthetic principles described by Graves, while the others violate, at least, one of these principles. Each slide is shown for approximately 45–60 seconds to the subject, who chooses one image per slide. The score of the test corresponds to the number of correct choices. All slides are in black, white and green. All images are abstract. The images of each slide are similar in style and in terms of the elements present. The average percentage of correct

[1] Photos of DJT can be found at: http://www.flickr.com/photos/robgiampietro/sets/72157611584992173/with/3136292750/.

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

Graves (1948) reported that art students achieved higher scores in the test than non-art students. He stated that: "the test's ability to differentiate the art groups from the non-art groups is unmistakably clear". Eysenck and Castle (1971) obtained different results showing fewer differences between art and non-art students (64.4 % vs. 60 %) with variances below 4 % in all cases, and also different responses in males and females. Eysenck and Castle (1971) pointed out the "general climate of art teaching, which now tends to stress simplicity and regularity to a greater extent than 25 years ago" as a possible reason for the differences observed. The DJT test was used as an instrument by the career advisors of the Portuguese Institute for Employment and Vocational Training. According to the results found by this institute while validating the test for the Portuguese population, published in internal reports and provided to the career advisors, the results achieved in the DJT with randomly selected individuals yield an average percentage of 50.76 % correct answers. This score is similar to the one obtained by answering randomly to the test, which indicates its difficulty. If we consider students in the last years of Fine Arts degrees, the average increases up to 61.87 %. Nevertheless, Götz and Götz (1974) report that "22 different arts experts (designers, painters, sculptors) had 0.92 agreement on choice of preferred design, albeit being critical of them" (Chamorro-Premuzic and Furnham 2004).

Like in most psychological tests, one should exercise great care when interpreting the results. The fact that a subject obtains a higher score in the DJT than another does not imply that he has better aesthetic judgement skills. It can mean, for instance, that one of the subjects is making choices based on aesthetics while the other is not. For example, a structural engineer may be inclined to choose well-balanced and stable designs, systematically valuing these properties above all else and ignoring rhythm, contrast, dynamism, etc. because the balance of the structure is the key factor to him. The test has been used for career guidance based on the reasoning that a subject that consistently makes choices according to aesthetic criteria is likely to have a vocation for an art-related career.

The DJT is based on aesthetic principles which may not be universally accepted or applicable (Eysenck 1969, Eysenck and Castle 1971, Uduehi 1995). Additionally, even if the aesthetic principles are accepted, the ability of the test to assess them has been questioned (Eysenck and Castle 1971). The average results obtained by humans in these tests also vary between studies (Eysenck and Castle 1971, Uduehi 1995). Although this can be, at least partially, explained by the selection of participants and other exogenous factors, it makes it harder to understand what constitutes a good score in this test.

The ability of these tests to measure the aesthetic judgement skills of the subjects is not undisputed, nor are the aesthetic principles they indirectly subscribe. Nevertheless, they can still be valuable validation tests in the sense that they can be used to measure the ability of an AJS to capture the aesthetic proprieties explored in these tests and the degree of accordance with the aesthetic judgements they implicitly defend.

11.2.2 User Evaluation and Popularity Prediction

The most obvious way of validating an AJS (at least one with learning capacities) may be to employ a set of images pre-evaluated by humans. The task of the AJS is to classify or "to assign an aesthetic value to a series of artworks which were previously evaluated by humans" (Romero et al. 2003).

There are several relevant papers published in the image processing and computer vision research literature that are aimed at the classification of images based on aesthetic evaluation. Most of them employed datasets obtained from photography websites. Some of those datasets are public, so they allow testing of other AJSs. In this section we perform a brief analysis of some of the most prominent works of this type.

Ke et al. (2006) proposed the task of distinguishing between "high quality professional photos" and "low quality snapshots". These categories were created based on users' evaluations of a photo website, so, to some extent, this can be considered as a classification based on aesthetic preference. The website was the dpchallenge.com photography portal, and they used the highest and lowest rated 10 % images from a set of 60,000 in terms of average evaluation. Each photo was rated by at least 100 users. Images with intermediate scores were not considered.

The authors employed a set of high-level image features (such as spatial distribution of edges, colour distribution, blur, hue count) and a support vector machine classification system, obtaining a correct classification rate of 72 %. Using a combination of these metrics with those published by Tong et al. (2004), Ke et al. (2006) achieved a success rate of 76 %.

Luo and Tang (2008) employed the same database. The 12,000 images of the dataset are accessible online[2] allowing the comparison of results. Unfortunately, neither the statistical information of the images (number of evaluations, average score, etc.) nor the images with intermediate ratings are available. The dataset is divided into two sets (training and test), made up of 6,000 images each. The authors state that these sets were randomly created. However, when one reverses the role of the test and training sets (i.e. training with original "test" set and testing with the original "training" set) the results differ significantly. This result indicates that the test and training set are not well-balanced.

Additionally, Luo and Tang (2008) used a blur filter to extract the background and the subject from each photo. Next, they employed a set of features related to clarity contrast (the difference between the crispness of the subject region and the background of the photo), lighting, simplicity, composition and colour harmony. They obtained a 93 % success rate using all features, which clearly improved upon previous results. The "clarity contrast" feature alone yields a success rate above 85 %. The authors pointed out that the difference between those results and the ones obtained by Ke et al. (2006) can be derived from the application of metrics to the image background regions and to the greater adequacy of the metrics itself.

[2] http://137.189.97.48/PhotoqualityEvaluation/download.html.

Datta et al. (2006) employed colour, texture, shape and composition, high-level ad-hoc features and a support vector machine to classify images gathered from a photography portal (photo.net). The dataset included 3581 images. All the images were evaluated by at least two persons. Unfortunately, the statistical information from each image, namely number of votes, value of each vote, etc. is not available. Similarly to previous approaches, they considered two image categories: the highest rated images (average aesthetic value ≥5.8, a total of 832 images) and the lowest rated ones (≤4.2, a total of 760 images), according to the ratings given by the users of the portal. Images with intermediate scores were discarded. Datta's justification for making this division is that photographs with an intermediate value "are not likely to have any distinguishing feature, and may merely be representing the noise in the whole peer-rating process" (Datta et al. 2006). The system obtained 70.12 % classification accuracy. The authors published the original dataset of this experiment, allowing future comparisons with other systems.

Wong and Low (2009) employed the same dataset, but selected the 10 % of the highest and lowest rated images. The authors extracted the salient regions of images, with a visual saliency model. They used global metrics related to sharpness, contrast, luminance, texture details, and low depth of field; and features of salient regions based on exposure, sharpness and texture details. Using a support vector machine classifier they obtained a 78 % 5-fold cross-validation accuracy.

In order to create a basis for research on aesthetic classification, Datta et al. (2008) proposed three types of aesthetic classification: aesthetic score prediction; aesthetic class prediction and emotion prediction. All the experiments explained in this section rely on aesthetic class prediction. He also published four datasets: the one employed in Datta et al. (2006), and 3 other extracted from photo.net (16,509 images), dpchallenge.com (14,494 images) and "Terragalleria" (14,494 images).[3] These three datasets include information regarding the number of votes per image and "score" (e.g. number of users that assigned a vote of "2" to image "id454"). Moreover, a dataset is included from the website "Alipr" with 13,100 emotion-tagged images.

Although not within the visual field, it is worth mentioning the work carried out by Manaris et al. (2007) in which a system was trained to distinguish between popular (high number of downloads) and unpopular classical music (low number of downloads). The dataset was obtained from downloads of the website Classical Music Archive (http://www.classicalarchives.com) in November 2003. Two sets, with high and low number of downloads, were created, in a similar way to the previously mentioned works. The "popular" set contained 305 pieces, each one with more than 250 hits, while the "not popular" contained 617 pieces with less than 22 downloads. The system is based on a set of metrics based on Zipf's Law applied to musical concepts such as pitch, duration, harmonic intervals, melodic intervals, harmonic consonance, etc. The classification system is based on an artificial neural network. The success rate was 87.85 % (it classified correctly 810 out of 922 instances), which

[3] Available from http://ritendra.weebly.com/aesthetics-datasets.html.

was considered promising by the authors. The same approach could be applied to images if we use the number of times an image is downloaded or the number of hits of its high-resolution version.

All these works rely on the use of photography and artistic websites. While these sites provides large datasets created by a third party, which should minimise the chances of being biased, the approach has several shortcomings for the purposes of AJS validation.

The experimental environment (participants and methodology) is not as controlled as in a psychological test, and several exogenous factors may influence the image scores. It is not possible to have all the information about the people and the circumstances in which they participated. The personal relations between users may affect their judgement. The same person may cast more than one vote, and so on.

It is also difficult to know what the users are evaluating when they vote. At photo.net the users can classify each image according to its "aesthetic" and "originality", however these scores are highly correlated (Datta et al. 2006), which indicates that users were not differentiating between these criteria. Since the selection of images is not under the control of the researcher, the aesthetic evaluation can be highly influenced by the semantics of content, novelty, originality and so on. These websites include some level of competition (in fact dpchallenge.com is a contest), so the possibilities of some biased votes is even higher.

The interpretation of the results obtained by an AJS in this kind of test is not straightforward. Different datasets have different levels of difficulty. As such, a percentage of correct answers of, e.g. 78 % can be a good or a bad score. As such, the comparison with the state of the art becomes of huge importance. Additionally, it may also be valuable to consider the difficulty of the task for humans. Thus, estimate the discrepancy between the success rate of the AJS and the success rates obtained by humans. Although this is not possible for the previously mentioned datasets, if the dataset includes all the voting information, one can calculate the agreement between humans and the AJSs. In other words, check if the response of the AJS is within the standard deviation for human responses.

For the purposes of AJS validation, the dataset should neither be trivial nor allow shortcuts that enable the system to perform the task exploiting properties of the artefacts which are not related with the task. Teller and Veloso (1996) discovered that their genetic programming approach to face recognition was identifying subjects based on the contents of the background of images (the photographs had been taken in different offices) instead of on the faces. The same type of effect may happen in aesthetic judgement test unless proper measures are taken. For instance, good photographers tend to have good cameras and take good photographs. A system may correctly classify photographs by recognising a good camera (e.g. a high resolution one) instead of recognising the aesthetic properties of the images. Thus, it is necessary to take the appropriate precautions to avoid this type of exploitation (e.g. reducing all the images to a common resolution before they are submitted to the classifier). This precaution has been taken in the works mentioned in Sect. 11.3 of this chapter. Nevertheless, it is almost impossible to ensure that the judgements are made exclusively on aesthetic properties.

For all the above reasons, the use of several datasets and types of tasks during the validation can help assessing the consistency and coherence of the results.

Creating datasets specifically for the purposes of the validation of AJSs is also valuable. An option is to create a dataset made up of images evaluated by humans in a controlled environment, following, for instance, a methodology similar to the one employed by Nadal (2007). We are not aware of any AJS evaluated like this in the field of visual art. In the musical field, there is a system that follows this approach (Manaris et al. 2005), in which a classifier is trained from human responses to musical pieces in a controlled experiment. A system similar to the one previously described achieved an average success rate of over 97 % in predicting (within one standard deviation) human emotional responses to those pieces (Manaris et al. 2007). Another option would be to create datasets that focus on a specific aesthetic property. For instance, to judge the balance of the composition one could ask photographers to take several pairs of photographs of the same motif, with the same camera, exposure, lighting conditions, etc. but with different framings so that one is a well-balanced composition and the other is not, according to the views of the photographers. This would allow the elimination of several of the external factors that could bias the judgement and would also allow an incremental development of the AJSs by focusing on one property at a time, and then moving towards tasks that require taking several aesthetic properties into consideration.

11.2.3 Style and Author Classification

In order to provide objective testing and to further analyse the abilities of AJSs, we explore validation approaches which test the ability of the system to learn the characteristics of a visual style (from an author, a trend, etc.). This type of test is not directly related with aesthetic value, but it can support AJS development.

In the field of computational creativity, a style-based classifier could allow the creation of image generation systems that produce images of a given artistic style and, perhaps more importantly in that context, it could be used to create images that are stylistically different from a given style or styles.

An objective way of performing this kind of test is employing artworks from several authors. The problems with this method usually arise from: (i) the relatively "low" production of most artists, since a machine learning approach can easily require hundreds or even thousands of examples; (ii) the heterogeneity of the artistic production of the authors, caused by the exploration of different styles, differences between early and mature works, etc. One can partially overcome these difficulties by selecting authors with vast productivity and by choosing the most prototypical works. Unfortunately, this may rule out the possibility of using several influential artists and bias the results by making the task easier than what would be desirable.

Another approach consists of classifying artworks according to the artistic "style". The main difficulties to overcome when setting up this type of experiment are: (i) the images must be previously, and correctly, classified as belonging to a particular style; (ii) one must ensure that there is no overlap between styles; (iii) one cannot use exclusively the most representative images of each style, otherwise the tasks may become trivial and, therefore, useless.

The first problem can be partially solved by using a relevant external source for the images. Unfortunately, the only published digital sets of artistic images we are aware of are those provided by Directmedia/The Yorck Project publications. However, the quality of the collections is far from perfect (they include black and white versions of some images, frames, detailed images of parts of other artworks, etc.). One can also resort to online databases of paintings. The collection "Oil paintings by Western masters" contains 46,000 images and can be found in the peer-to-peer network. The *Worldimages* website (http://worldimages.sjsu.edu/kiosk/artstyles.htm), the website http://www.zeno.org, developed by the creators of "The Yorck Project", and online museum websites are also good sources of images.

Wallraven et al. (2008) analysed the perceptual foundations of the traditional categorisation of images into art styles, finding supporting evidence. They concluded that style identification was predominantly a vision problem and not merely a historical or cultural artefact.

Wallraven et al. (2009) presented an experiment that analysed the capacity of a group of non-experts in art to categorise a set of artworks in styles. One of the metrics they analysed is the artist consistency, which was higher if paintings of the same painter were put in the same cluster. In one experiment, they obtained an average artist consistency of 0.65. The conclusions were that "experts were able to reliably group unfamiliar paintings of many artists into meaningful categories". In the same paper, the authors employed a set of low-level measures (Fourier analysis, colour features, Gist, etc.) and a k-means algorithm to categorise the artworks into styles. They concluded that low-level features were not adequate to artistic style classification: "the fact that neither texture, nor colour-based, scale-sensitive or complexity measures correlate at any dimension casts doubt on whether another [low level] measure will do much better" (Wallraven et al. 2008).

Marchenko et al. (2005), based on the colour theory of Itten (1973), characterised regions of the image in terms of "artistic colour concepts", while Yan and Jin (2005) used several colour spaces to gather information with the aim of retrieving and classifying oil paintings.

There are several papers in the content-based image retrieval literature that propose image classification based on the "type" of image, distinguishing professional photos from amateur ones, e.g. (Tong et al. 2004); or photos from: (i) paintings (Cutzu et al. 2003), (ii) computer graphics (Athitsos et al. 1997), (iii) computer-generated images (Lyu and Farid 2005). These tasks result in an interesting test field for AJS, creating the opportunity of using AJSs in image classification tasks that are far from aesthetics. These works can also provide tools (e.g., features, classification methods, etc.) of interest to the creative computer community, in particular to those researchers involved in artistic tasks.

11.3 The Evolution of an AJS

This section describes the evolution of an AJS over the course of the past decade. It started as a heuristic based system, it was tested using the DJT, and it subsequently became part of an evolutionary art tool. Prompted by the results obtained, an AJS with learning abilities was developed and tested in a wide variety of experiments, which are also described briefly.

11.3.1 A Heuristic AJS

Machado and Cardoso (1998) took inspiration from the works of Arnheim (1956; 1966; 1969), as well as from the research indicating a preference for simple representations of the world, and a trend to perceive it in terms of regular, symmetric and constant shapes (Wertheimer 1939, Arnheim 1966, Tyler 2002, Field et al. 2000). They explored the working hypothesis that the aesthetic value was linked with the sensorial and intellectual pleasure experienced when finding a compact percept (i.e. internal representation) of a complex visual stimulus (cf. Chap. 12). The identification of symmetry, repetition, rhythm, balance, etc. can be a way of reducing the complexity of the percept, which would explain the universal nature of these aesthetic principles and the ability of the brain to recognise them "effortlessly".

The approach rewards images that are simultaneously visually complex and easy to perceive, employing estimates for the *Complexity of the Percept* (CP) and for the *Complexity of the Visual Stimulus* (CV). An estimate for CV should assess the predictability of the image pixels. JPEG image compression mainly affects the high frequencies, which can normally be discarded without significant loss in image quality. The amount, and quality (i.e. the error involved) of the compression achieved by this method depends on the predictability of the pixels in the image being compressed. Unlike JPEG compression, which only takes into account local information, fractal image compression can take advantage of the self-similarities present in the image. Machado and Cardoso (1998) assume that JPEG compression is less like the way humans perceive images than fractal image compression, and hence use fractal compression as a rough estimate of the CP. CP and CV are estimated through the division of the root mean square error by the compression ratio resulting, respectively, from the fractal (quadratic tree based) and JPEG encoding of the image.

A time component is also considered (Machado and Cardoso 1998; 2002). As time elapses, there is a variation in the detail level of image perception. Therefore, it is necessary to estimate CP for specific points in time, in this case t_0 and t_1, which is achieved by carrying out a fractal image compression with increasing detail levels. The proposed approach values images where CP is stable for different detail levels. The idea being that as time goes by one should be able to acquire additional information about the image, for example: the increase in size of the percept should be balanced out by the increase in its level of detail. It is important to notice that Machado and Cardoso neither suggested that the employed JPEG complexity was able to fully capture the concept of image complexity, nor that the fractal image compression was able to capture the complexity of visual perception. They posited that JPEG was closer to visual complexity than fractal compression, and that fractal compression was closer to processing complexity than JPEG, subsequently testing the possibility of using these measures as rough estimates for these concepts in the context of a specific, and limited, aesthetic theory.

The following formula was proposed as a way to capture the previously-mentioned notions (Machado and Cardoso 1998):

$$\text{aesthetic value} = \frac{CV^a}{(CP(t_1) \times CP(t_0))^b} \times \frac{1}{\left(\frac{CP(t_1)-CP(t_0)}{CP(t_1)}\right)^c} \tag{11.1}$$

where a, b and c, are parameters used to tune the relevance given to each of the components. The left side of the formula rewards those images which have high CV and low CP estimates at the same time, while the right side rewards those images with a stable CP across time. The division by $CP(t_1)$ is a normalisation operation. The formula can be expanded in order to encompass further instants in time, but the limitations of the computational implementation led the authors to use only two instants in their tests.

The images of the DJT were digitalised, converted to greyscale, and resized to a standard dimension of 512×512 pixels, which may involve changes in the aspect ratio. The estimates for CV, $CP(t_1)$ and $CP(t_0)$ were computed for the resulting images. Using these estimates, the outcome of formula (11.1) was calculated for each of the images. For each of the 90 pairs or triads of images comprising the DJT, the system chose the image that yielded a higher value according to formula (11.1).

The percentage of correct answers obtained by the AJS depends on the values of the parameters a, b and c. Considering all combinations of values for these parameters ranging in the $[0.5, 2]$ interval with 0.1 increments, the maximum percentage of correct answers was 73.3 % and the minimum 54.4 %. The average success rate of the system over the considered parametric interval was 64.9 %.

As previously mentioned, the highest average percentage of correct answers in human tests in the DJT reported by Eysenck and Castle (1971) is 64.4 %, and was obtained by subjects that were final year fine art graduates, a value that is surprisingly similar to the average success rate of our system (64.9 %).

Although comparing the performance of the system to the performance of humans is tempting, one should not jump to conclusions! A similar result cannot be interpreted as a similar ability to perform aesthetic judgements. As previously mentioned, humans may follow principles that are not exclusively in aesthetic order to choose images. Moreover, since the test aims at differentiating between humans, it may take for granted principles that are consensual between them, and the AJS would be unable to identify. Finally, the results say nothing regarding the validity of the test itself (a question that is outside the scope of our research). Thus, what can be concluded is that the considered formulae and estimates are able to capture some of the principles required to obtain a result that is statistically different from the one obtained by answering randomly in the DJT.

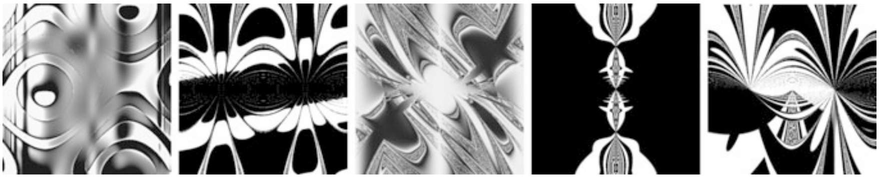

Fig. 11.1 Examples of images created using an Evolutionary Engine and heuristic AJS

Some constraints were applied to the different formula components so as to explore these ideas in an evolutionary context, in the following way:

$$\mathit{fitness} = \frac{\min(\alpha, CV)^a}{\max(\beta, CP(t_1) \times CP(t_0))^b} \times \frac{1}{\max\bigl(\gamma, \frac{CP(t_1)-CP(t_0)}{CP(t_1)}\bigr)^c} \qquad (11.2)$$

where α, β and γ are constants defined by the user.

These constraints are necessary to ensure that the evolutionary algorithm does not focus exclusively on one of the components of the formula. This could make it converge to images with maximum visual complexity (e.g. white noise images) disregarding entirely the processing complexity estimates, or to images with minimal processing complexity estimates (e.g. pure white). It was not necessary to make additional changes to prevent the situation where $CP(t_1) \simeq 0$ because these images have very low fitness, and are, therefore, already avoided by the evolutionary algorithm.

It is important to notice that the situations where $CP(t_1) \simeq 0$ or $CP(t_1) - CP(t_0) \simeq 0$, although theoretically possible, never occurred when using natural imagery.

Machado and Cardoso (2002) carried out various experiments using a Genetic Programming engine and formula (11.2) as the fitness function.

The results achieved with this autonomous evolutionary art system are quite striking (Machado and Cardoso 2002). In spite of the shortcomings—e.g. it only deals with greyscale images—it allows the evolution of a wide variety of images with different aesthetic merits. Figure 11.1 shows the fittest images from several independent runs.

11.3.2 Learning AJSs

Based on the results described in the previous section, we developed a learning AJS. The system consists of two modules: a *Feature Extractor* (FE) and an *adaptive classifier*.

The FE performs an analysis of the input images by collecting a series of low-level feature values, most of which are related to image complexity. The values that result from the feature extractor are normalised between 1 and −1. These values are the inputs of the classifier, which is made up of a feed-forward artificial neural network with one hidden layer. For training purposes, we resorted to SNNS (*Stuttgart*

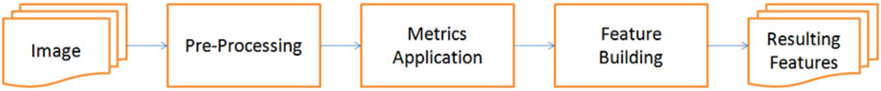

Fig. 11.2 Feature extraction steps

Neural Network Simulator, Zell et al. 2003) and standard back-propagation. The results presented in this chapter concern artificial neural networks with one input unit per feature, 12 units in the hidden layer, and 2 units in the output layer (one for each category). A training pattern specifying an output of (1; 0) indicates that the corresponding image belongs to the first set. Likewise, a training pattern with an output of (0; 1) indicates that the corresponding image belongs to the second set. The parameters for the classifier and FE were established empirically in previous experiments.

The experiments presented in this section concern classification tasks of different nature: aesthetic value prediction, author identification and popularity prediction. All the results presented in this section were obtained by the same AJS, trained in different ways. Finally, we describe the integration of this AJS with an evolutionary image generation system.

11.3.2.1 Feature Extraction

In this section we describe the feature extraction process.

The feature extraction can be summarised to the following steps (see Fig. 11.2): (i) *Pre-processing,* which includes all the transformation and normalisation operations applied to a given input image; (ii) *Metrics application*, that is, the application of certain methods based on statistical measurements and image complexity estimates; (iii) *Feature building*, the extraction of results from the metrics applied in order to build the image feature set.

Pre-processing The images from a dataset are individually submitted to a series of transformations before being analysed. A given input image is loaded and resized to a standard width and height of 256×256 pixels, transformed into a three-channel image in the RGB (red, green and blue) colour space, with a depth of 8 bits per channel and all the pixel values are scaled to the $[0, 255]$ interval. This step ensures that all input images share the same format and dimensions.

Next, the image is converted into the HSV (Hue, Saturation and Value) colour space and its HSV channels are split. Each of these channels is stored as a one-channel greyscale image. From here on, we will refer to these images as H, S and V channel images. A new greyscale image is also created by performing a pixel by pixel multiplication of S and V channels and scaling the result to $[0, 255]$. From now on, we will refer to this image as the CS (Colourfulness) channel image.

The images resulting from these operations are subject to transformation operations. The current version of the FE supports seven transformations: no filter,

Table 11.1 Fractal image compression parameters

	Low	Medium	High
Image size	256 × 256 pixels		
Minimum partition level	2	2	3
Maximum partition level	4	5	6
Maximum error per pixel	8	8	8

which means no transformation applied; Sobel-based (Sobel 1990) and Canny-based (Canny 1986) edge detection of horizontal and vertical edges, horizontal edges, vertical edges.

Metrics Application A set of metrics is applied to the images resulting from the pre-processing operations. The FE calculates the following metrics: average (i) and standard deviation (ii) of the image pixel values; complexity estimates based on JPEG (iii) and fractal compression (iv); Zipf Rank-Frequency (v) and Size-Frequency (vi), which result from the application of the Zipf's law (Zipf 1949); (vii) Fractal dimension estimates using the box-counting method (Taylor et al. 1999).

The average (i) and standard deviation (ii) are calculated using the pixel intensity value of each image, except for the H channel image. Since the Hue channel is circular, the average and the standard deviation are calculated based on the norm and angle of Hue values. In addition, a multiplication of the Hue angle value by the CS value is made and consequently a norm is calculated using Hue and CS values.

The image compression schemes used are lossy and so there will be compression errors, i.e. the compressed image will not exactly match the original. All other factors being equal, complex images will tend toward higher compression errors and simple images will tend toward lower compression errors Additionally, complex images will tend to generate larger files than simple ones. Thus, compression error and file size are positively correlated with image complexity.

We consider three levels of detail for the JPEG (iii) and Fractal compression (iv) metrics: low, medium, and high. For each compression level the process is the same, the image is encoded in JPEG and fractal format. In the experiments described herein, we use a quad-tree fractal image compression scheme (Fisher 1995) with the set of parameters given in Table 11.1.

The calculation of the Zipf Rank Frequency (v) metrics implies: counting the number of occurrences of each pixel intensity value in the image; ordering them according to the number of occurrences; tracing a rank vs. number of occurrences plot using a logarithmic scale in both axis; calculating the slope of the trendline and the linear correlation with the trendline.

For the Hue channel, this metrics is calculated in two ways: (i) as described above; (ii) instead of counting the number of occurrences of each Hue value, we add the CS channel values of the corresponding pixels (and divide them by 255 for normalisation purposes). The rationing is that the perceived Hue depends on the saturation and value of the corresponding pixel.

The Zipf Size Frequency (vi) metric is calculated in similar way to Zipf Rank Frequency. For each pixel we calculate the difference between its value and each of its neighbouring pixels. We count the total number occurrences of differences in size 1, size 2, …, size 255. We trace a size vs. number of occurrences plot using a logarithmic scale in both axes and we calculate the slope and linear correlation of the trendline.

For the H channel we consider a circular distance. The Hue Size Frequency is also calculated using the CS channel. The last metric is a Fractal Dimension estimate (vii) based on the box-counting method. Briefly described: the box-counting method computes the number of cells (boxes) required to cover an object entirely, with grids of cells of varying box size.

Feature Building After the application of the metrics, the results are aggregated to make up the image features.

The average and standard deviation for each channel image returns two values, except for the Hue channel that returns four values for the average and two values for the standard deviation. The JPEG and Fractal compression metrics return three values each, corresponding to the three compression levels considered. Although these metrics are applied to all the images resulting from the pre-processing transformations, the JPEG metric is also applied to the RGB image. As for the Zipf's law based metrics and fractal dimension, the slope of the trendline (m) and the linear correlation (R2) of all greyscale images are extracted. In the case of the Hue channel, these metrics return four values each: two considering only the Hue channel and two considering the Hue and CS channel. We employ a total of 53 metrics applied to seven pre-processing operators, which yield 371 features per image.

11.3.2.2 DJT Experiments

The main goals of these experiments were: (i) confirming the results described in the previous section by the heuristic based AJS and (ii) determining the viability of training an artificial neural network for aesthetic judgement tasks from a small set of examples.

We train an artificial neural network using some of the DJT items and test its ability to predict the correct choice on the remaining ones. The network receives as input the features of two images from the same slide. The output indicates the chosen one. Each of the 82 DJT items that consist of two images yields a "pattern". Eight of the 90 DJT items contain three images instead of two. To deal with these cases, each of these eight items was divided into two "patterns", using the "correct" image in both patterns. Thus, each triad results in two patterns, which yields a total number of 98 patterns (82 obtained from pairs and 16 from triads).

Due to the small number of training patterns we employed a 20-fold cross-validation technique. 20 sets were created from the 98 patterns (18 with 5 patterns and 2 with 4 patterns). In each of the 20 "folds", 19 of the sets were used for training while the remaining one was used for validation.

The sets were generated at random and care was taken to ensure that the two patterns resulting from an item with three images were integrated into the same set. Thus, it was guaranteed that the correct image was not simultaneously used for training and testing the neural network.

Considering the 20 experiments carried out, the global success rate in the test sets was 74.49 %. Which corresponds to a percentage of 71.67 % correct answers in the Design Judgment Test.[4] The result is similar to the maximum success rate previously achieved with the heuristic AJS (73.3 %) by adjusting the parameters. This reinforces the conclusion that it is possible to capture some of the aesthetic principles considered by Maitland Graves in the DJT. They also show that it is possible to learn principles of aesthetic order based on a relatively small set of examples. The fact that the approach was not able to achieve the maximum score in the DJT has two, non exclusive, explanations: (i) the features are unable to capture some of the aesthetic principles required to obtain a maximum score in the DJT; (ii) the set of training examples is not sufficient to allow the correct learning of these principles.

Although the results obtained by the system are higher than the human averages reported in the previously mentioned studies, these results are not comparable. In addition to the issues we mentioned when analysing the results of the heuristic based classifier, the nature of the task is different herein: humans do not make their choices based on a list of correct choices for other items of the test.

11.3.2.3 Author Identification Experiments

In Machado et al. (2004) we presented the results obtained by a previous version of our AJS in an author identification task. The image dataset was made up of 98 paintings from Goya, 153 from Monet, 93 from Gauguin, 122 from Van Gogh, 81 from Kandinsky, and 255 from Picasso. Although the system obtained high success rates (above 90 %), further experiments revealed that the reduced number of images and their nature made the classification task easier than expected.

Taking into account the dataset limitations mentioned in Sect. 11.2.2, we created a dataset composed of images from three prolific painters, from chronologically consecutive artistic movements:

Claude-Oscar Monet (Impressionism, mid 19th century). It consists of 336 images, most of them landscapes and portraits.

Vincent van Gogh (Post-Impressionism, late 19th century): a total number of 1046 well-known images from his work, including landscapes, portraits, self-portraits and still lifes.

Pablo Picasso (Cubism and Surrealism, early 20th century): a total of 540 images belonging to different stages were used, ranging from the Blue Period to the author's surrealist stage.

We avoided using greyscale images and images with insufficient resolution. Some of the images (12 from Picasso and 8 from Van Gogh) included the frames

[4] Some of the test items are triads, hence the lower percentage.

Table 11.2 Success rate in validation set (the results are averages of 50 independent runs)

Picasso vs. Van Gogh	Picasso vs. Monet	Van Gogh vs. Monet
92.1 %	91.5 %	89.9 %

Table 11.3 Confusion matrix (the results are averages of 50 independent runs)

	Picasso	Monet	Van Gogh
Picasso	87.59 %	2.59 %	9.81 %
Monet	4.76 %	70.24 %	25.00 %
Van Gogh	4.11 %	6.60 %	89.29 %

of the painting. Since we avoided doing any sort of manual pre-processing of the images, the frames were not removed. The images were gathered from different sources and the dataset will be made available for research purposes, thus enabling other researchers to compare their results with ours.

The experimental results are averages of 50 independent runs using different training and validation sets. In each run, 90 % of the images were randomly selected to train the artificial neural network. The remaining ones were used as validation set to assess the performance of the artificial neural network. The training of the artificial neural network was stopped after a predetermined number of learning steps. All the results presented concern the performance in validation.

Table 11.2 presents the results obtained in an author classification task with two classes. As it can be observed, discriminating between the works of Van Gogh and Monet was the biggest challenge. Conversely, Pablo Picasso's works were easily distinguished from the ones made by Monet and Van Gogh.

In Table 11.3 we present the confusion matrix for this experiment, which reinforces the previous findings. There is a significant drop in performance when it comes to the correct identification of Claude-Oscar Monet's works. The existence of fewer paintings of this author can explain the difficulties encountered in correctly learning how to recognise his style. A more detailed analysis of this experiment is currently in preparation.

Overall, the results indicate that the considered set of metrics and classifier system are able to distinguish between the signatures (in the sense used by Cope 1992) of different authors. It cannot be stated that the AJS is basing its judgement, at least exclusively, on aesthetic principles. It can, however, be stated that it is able to perform stylistic classification in the considered experimental settings. Even if we could demonstrate that the system was following aesthetic principles, this would not ensure that those principles are enough to perform aesthetic value assessments. If the system obtained bad results in distinguishing between works that have different aesthetic properties it would cast serious doubts on its ability to perform aesthetic evaluation. Thus, a good performance on an author identification task does not ensure the ability to perform aesthetic evaluation, but it is arguably a prerequisite.

11.3.2.4 Image Classification Based on Online Evaluation

We used the dataset provided by Datta et al. (2006) that was analysed in Sect. 11.2.2. The database contains 832 images with an aesthetic rating ≥5.8 and 760 images with a rating ≤4.2. However, when we carried out our experiment, some of the images used by Datta were no longer available at photo.net, which means that our image set is slightly smaller. We were able to download 656 images with a rating of 4.2 or less, and 757 images with a rating of 5.8 or more.

We conducted 50 runs, each with different training and validation sets, randomly created with 80 % and 20 % of the images, respectively. The success rate in the validation set was 77.22 %, which was higher than the ones reported in the original paper (Datta et al. 2006) but lower than the one obtained by Wong and Low (2009), using 10 % of the images in each set.

11.3.2.5 Integration in an Image Generation System

A previous version of the AJS described here was used in conjunction with a genetic programming evolutionary art tool. The main goal of this experiment, reported by Machado et al. (2007), was to develop an approach that promoted stylistic change from one evolutionary run to the next. The AJS assigns fitness to the evolved images, guiding the evolutionary engine.

The AJS is trained by exposing it to a set of positive examples made up of artworks of famous artists, and to a set of negative examples made up of images generated randomly by the system. The goal is twofold: (i) evolving images that relate with the aesthetic reference provided by the positive examples, which can be considered an inspiring set; (ii) evolving images that are novel relative to the imagery typically produced by the system. Thus, more than trying to replicate a given style, the goal is to break from the traditional style of the evolutionary art tool. Once novel imagery is found (i.e. when the evolutionary engine is able to find images that the AJS fails to classify as being created by it), these images are added to the negative set of examples, the AJS is re-trained and a new evolutionary run begins. This process is iteratively repeated and, by this means, a permanent search for novelty and deviation from the previously explored paths is enforced.

Next, the genetic programming engine and the AJS performed 11 consecutive iterations (Machado et al. 2007). In each iteration, the evolutionary engine was able to find images that were misclassified by the AJS. Adding this set of examples to the dataset forced the AJS to find new ways to discriminate between paintings and the images created by the evolutionary art tool. The evolutionary engine and the AJS performed well across all iterations. The success rate of the AJS for validation set images was above 98 % in all iterations. The evolutionary engine was also always able to find novel styles that provoked misclassification errors. In Fig. 11.3 we present some examples of images created in the 1st and 11th iteration.

Overall, the results indicate that the internal coherency of each run is high, in the sense that runs converge to imagery of a distinctive and uniform style. The style

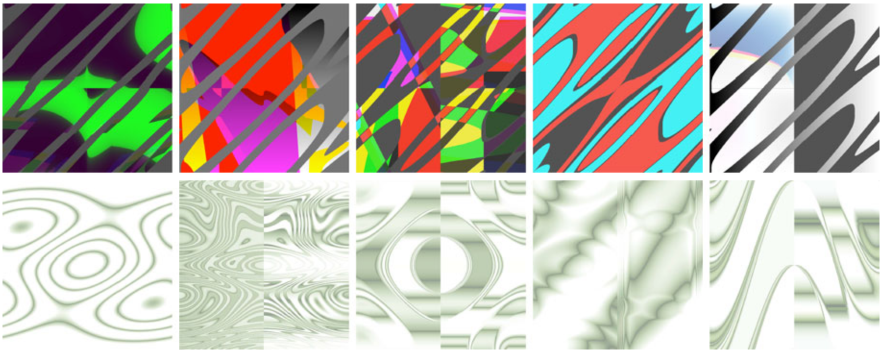

Fig. 11.3 Examples of images created using an Evolutionary Engine and an adaptive AJS in the 1st (*upper row*) and 11th (*lower row*) iteration of the experiment

Table 11.4 Percentage of images classified as external by the ANNs used to guide evolution in iterations 1 and 11, and the difference between them

Set	Iteration 1	Iteration 11	Difference
Painting masterpieces	99.68 %	96.88 %	−2.80 %
User-guided evolution	17.99 %	10.07 %	−7.91 %

differences between runs are also clear, indicating the ability of the approach to promote a search for novelty. They also indicate that the aesthetic reference provided by the external set manages to fulfil its goal, making it possible for AJSs to differentiate between those images that may be classified as paintings and those generated by the GP system (Machado et al. 2007).

A set of experiments was carried out to compare the performance of the AJS from the 1st and 11th iteration, using datasets made up of images that were not employed in the runs. The experimental results are presented in Table 11.4 and show that the AJS of the 11th generation performs worse than the one of the 1st iteration at classifying external imagery (a difference of 2.8 %), and better at classifying evolution generated images (a difference of 7.91 %). These results suggest that the iterations performed with the evolutionary engine promote the generalisation abilities of the AJS, leading to an overall improvement in classification performance.

The integration of an AJS within a bootstrapping evolutionary system of this kind is extremely valuable. As the results indicate, it allows the generation of images that explore the potential weaknesses of the classifier system and the subsequent use of these images as training instances, leading to an overall increase in performance. Additionally, if the evolutionary system is able to generate images that the AJS is unable to classify correctly (even after re-training it) and that a human can classify, it shows that the set of features is not sufficient for the task at hand. Additionally, it gives indications about the type of analysis that should be added in order to improve the performance of the AJS.

11.4 Conclusions

The development of AJS presents numerous difficulties, and there are still several open questions, validation being one of them.

This chapter proposed several ways of testing and comparing the results of aesthetic judgement systems. We proposed validation tasks based on psychological tests, on style and author identification, on users' preferences, and on popularity prediction.

Some alternatives for AJS design have been briefly explored. We focus on an adaptive architecture based on a series of metrics and a machine learning classifier. This type of approach was employed in the field of computational creativity and is popular in content based image retrieval and computer vision research. Some of the works in these areas that can be valuable to computational creativity are analysed. The datasets and results they obtained are presented to serve as a reference for future comparison.

We also presented a heuristic based AJS and discussed the results obtained by the system in a psychological test designed for humans. The experiments show that this AJS was able to capture some of the aesthetic principles explored in the test. The integration of the heuristic AJS with an image generation system was also described and the results briefly discussed.

Subsequently, we described the development of an adaptive AJS based on complexity metrics and an artificial neural network classifier, and presented the experimental results obtained by this AJS in several validation tasks.

The results attained in the psychological test show that the system is able to learn from a set of examples made up of items of the test, obtaining a success rate above 70 % in a cross validation experiment. This result is similar to the one obtained by the heuristic based AJS, indicating that the system is able to reverse engineer some of the aesthetic principles considered in the DJT.

The author identification tasks show that, in the considered experimental settings, the system is able to perform classification based on the image style with an average success rate above 90 % in binary classification. The results obtained by our system in the prediction of users' aesthetic evaluation of online photographs are comparable with those reported as state of the art.

Finally, we presented the integration of the learning AJS with an image generation engine to build a system designed to promote a constant search for novelty and stylistic change.

Submitting the same AJS to several validation tasks allows one to overcome, at least partially, the shortcomings of individual tasks and to get additional insight on the weaknesses and strengths of the AJS.

We consider that the adoption of common validation procedures is an important step towards the development of the field. Sharing datasets allows other researchers to assess the strengths and weaknesses of their systems relative to published work. Sharing the training and test patterns used in experiments further promotes this collaboration between research teams, since it enables assessment of performance improvement that can be expected by the inclusion of the metrics used by other researchers in one's own AJS. Once these performance improvements are identified, the logical next step is the development, through collaboration, of AJSs that encompass the metrics used by the different research groups. These could lead, for instance, to an international research project where several research groups build a common AJS. Some of the groups could propose metrics, others design the classifier, and so on. Using the validation approaches proposed in this chapter (and future research in this area) it becomes possible to validate the classifier and compare the results with previous approaches. Moreover, due to the numerical nature of the validation approach, it is possible to identify relevant metrics in the classifier for the tasks considered.

AJSs can be valuable for real life applications, including:

- **Image Classification**—e.g., discriminating between professional and amateur photos, paintings and photos, images that are interesting to a particular user, etc.
- **Image Search Engines**—which could take into account user preference, or stylistic similarity to a reference image or images.
- **Online Shopping**—the ability to recognise the aesthetic taste of the user could be explored to propose products or even to guide product design and development.

The development of AJSs can also play an important role in the study of aesthetics, in the sense that the ability to capture aesthetic preferences of individuals and groups may promote a better understanding of the phenomena influencing aesthetic preferences, including cultural differences, training, education, trends, etc.

More importantly, the creation of systems able to perform aesthetic judgements may prove vital for the development of computational creativity systems. For instance, the development of an AJS that closely matches the aesthetic preferences of an individual would open a wide range of creative opportunities. One could use such an AJS in conjunction with an image generation system to create custom made "artificial artists" that would be able to create artworks which specifically address the aesthetic needs of a particular person. These systems could change through time, accompanying the development of the aesthetic preferences of the individual and promoting this development. They could also be shared between people as a way of conveying personal aesthetics, or could be trained to match the aesthetic preferences of a community in order to capture commonality. These are vital steps to accomplish our long term goal and dream: the development of computational systems able to create and feel their art and music.

Acknowledgements The authors would like to thank the anonymous reviewers for their constructive comments, suggestions and criticisms. This research is partially funded by: the Spanish Ministry for Science and Technology, research project TIN2008-06562/TIN; the Portuguese Foundation for Science and Technology, research project PTDC/EIA-EIA/115667/2009; Xunta de Galicia, research project XUGA-PGIDIT10TIC105008-PR.

References

Arnheim, R. (1956). *Art and visual perception, a psychology of the creative eye*. London: Faber and Faber.

Arnheim, R. (1966). *Towards a psychology of art/entropy and art—an essay on disorder and order*. The Regents of the University of California.

Arnheim, R. (1969). *Visual thinking*. Berkeley: University of California Press.

Athitsos, V., Swain, M. J., & Frankel, C. (1997). Distinguishing photographs and graphics on the world wide web. In *Proceedings of the 1997 workshop on content-based access of image and video libraries (CBAIVL '97), CAIVL '97* (pp. 10–17). Washington: IEEE Computer Society. http://portal.acm.org/citation.cfm?id=523204.791698.

Baluja, S., Pomerlau, D., & Todd, J. (1994). Towards automated artificial evolution for computer-generated images. *Connection Science*, *6*(2), 325–354.

Boden, M. A. (1990). *The creative mind: myths and mechanisms*. New York: Basic Books.

Burt, C. (1933). The psychology of art. In *How the mind works*. London: Allen and Unwin.

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *8*(6), 679–698.

Chamorro-Premuzic, T., & Furnham, A. (2004). Art judgement: a measure related to both personality and intelligence? *Imagination, Cognition and Personality*, *24*, 3–25.

Cope, D. (1992). On the algorithmic representation of musical style. In O. Laske (Ed.), *Understanding music with AI: perspectives on music cognition* (pp. 354–363). Cambridge: MIT Press.

Cutzu, F., Hammoud, R. I., & Leykin, A. (2003). Estimating the photorealism of images: distinguishing paintings from photographs. In *CVPR (2)* (pp. 305–312). Washington: IEEE Computer Society.

Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2006). Studying aesthetics in photographic images using a computational approach. In *Lecture notes in computer science*. *Computer vision—ECCV 2006, 9th European conference on computer vision, part III*, Graz, Austria (pp. 288–301). Berlin: Springer.

Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2008). Image retrieval: ideas, influences, and trends of the new age. *ACM Computing Surveys*, *40*, 5:1–5:60. http://doi.acm.org/10.1145/1348246.1348248.

Dorin, A., & Korb, K. B. (2009). Improbable creativity. In M. Boden, M. D'Inverno, & J. McCormack (Eds.), *Dagstuhl seminar proceedings: Vol. 09291*. *Computational creativity: an interdisciplinary approach*, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany. http://drops.dagstuhl.de/opus/volltexte/2009/2214.

Eysenck, H. (1969). Factor analytic study of the Maitland Graves Design Judgement Test. *Perceptual and Motor Skills*, *24*, 13–14.

Eysenck, H. J. (1983). A new measure of 'good taste' in visual art. *Leonardo, Special Issue: Psychology and the Arts*, *16*(3), 229–231. http://www.jstor.org/stable/1574921.

Eysenck, H. J., & Castle, M. (1971). Comparative study of artists and nonartists on the Maitland Graves Design Judgment Test. *Journal of Applied Psychology*, *55*(4), 389–392.

Eysenck, H. J., Götz, K. O., Long, H. Y., Nias, D. K. B., & Ross, M. (1984). A new visual aesthetic sensitivity test—IV. Cross-cultural comparisons between a Chinese sample from Singapore and an English sample. *Personality and Individual Differences*, *5*(5), 599–600. http://www.sciencedirect.com/science/article/B6V9F-45WYSPS-1M/2/1b43c2e7ad32ef89313f193d3358b441.

Field, D. J., Hayes, A., & Hess, R. F. (2000). The roles of polarity and symmetry in the perceptual grouping of contour fragments. *Spatial Vision*, *13*(1), 51–66.

Fisher, Y. (Ed.) (1995). *Fractal image compression: theory and application*. London: Springer.

Frois, J., & Eysenck, H. J. (1995). The visual aesthetic sensitivity test applied to Portuguese children and fine arts students. *Creativity Research Journal*, *8*(3), 277–284. http://www.leaonline.com/doi/abs/10.1207/s15326934crj0803_6.

Furnham, A., & Walker, J. (2001). The influence of personality traits, previous experience of art, and demographic variables on artistic preference. *Personality and Individual Differences*, *31*(6), 997–1017. http://www.sciencedirect.com/science/article/B6V9F-440BD9B-J/2/c107a7e1db8199da25fb754780a7d220.

Götz, K. (1985). *VAST: visual aesthetic sensitivity test*. Dusseldorf: Concept Verlag.

Götz, K. O., & Götz, K. (1974). The Maitland Graves Design Judgement Test judged by 22 experts. *Perceptual and Motor Skills*, *39*, 261–262.

Graves, M. (1946). *Design judgement test*. New York: The Psychological Corporation.

Graves, M. (1948). *Design judgement test, manual*. New York: The Psychological Corporation.

Graves, M. (1951). *The art of color and design*. New York: McGraw-Hill.

Itten, J. (1973). *The art of color: the subjective experience and objective rationale of color*. New York: Wiley.

Iwawaki, S., Eysenck, H. J., & Götz, K. O. (1979). A new visual aesthetic sensitivity test (vast): II. Cross cultural comparison between England and Japan. *Perceptual and Motor Skills*, *49*(3), 859–862. http://www.biomedsearch.com/nih/new-Visual-Aesthetic-Sensitivity-Test/530787.html.

Ke, Y., Tang, X., & Jing, F. (2006). The design of high-level features for photo quality assessment. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference*, *1*, 419–426.

Kowaliw, T., Dorin, A., & McCormack, J. (2009). An empirical exploration of a definition of creative novelty for generative art. In K. B. Korb, M. Randall & T. Hendtlass (Eds.), *Lecture notes in computer science: Vol. 5865. ACAL* (pp. 1–10). Berlin: Springer.

Luo, Y., & Tang, X. (2008). Photo and video quality evaluation: focusing on the subject. In D. A. Forsyth, P. H. S. Torr & A. Zisserman (Eds.), *Lecture notes in computer science: Vol. 5304. ECCV (3)* (pp. 386–399). Berlin: Springer.

Lyu, S., & Farid, H. (2005). How realistic is photorealistic? *IEEE Transactions on Signal Processing*, *53*(2), 845–850.

Machado, P., & Cardoso, A. (1998). Computing aesthetics. In F. Oliveira (Ed.), *Lecture notes in computer science: Vol. 1515. Proceedings of the XIVth Brazilian symposium on artificial intelligence: advances in artificial intelligence*, Porto Alegre, Brazil (pp. 219–229). Berlin: Springer.

Machado, P., & Cardoso, A. (2002). All the truth about NEvAr. *Applied Intelligence, Special Issue on Creative Systems*, *16*(2), 101–119.

Machado, P., Romero, J., & Manaris, B. (2007). Experiments in computational aesthetics: an iterative approach to stylistic change in evolutionary art. In J. Romero & P. Machado (Eds.), *The art of artificial evolution: a handbook on evolutionary art and music* (pp. 381–415). Berlin: Springer.

Machado, P., Romero, J., Manaris, B., Santos, A., & Cardoso, A. (2003). Power to the critics—a framework for the development of artificial art critics. In *IJCAI 2003 workshop on creative systems*, Acapulco, Mexico.

Machado, P., Romero, J., Santos, A., Cardoso, A., & Manaris, B. (2004). Adaptive critics for evolutionary artists. In R. Günther et al. (Eds.), *Lecture notes in computer science: Vol. 3005. Applications of evolutionary computing, EvoWorkshops 2004: EvoBIO, EvoCOMNET, EvoHOT, EvoIASP, EvoMUSART, EvoSTOC*, Coimbra, Portugal (pp. 435–444). Berlin: Springer.

Manaris, B., Romero, J., Machado, P., Krehbiel, D., Hirzel, T., Pharr, W., & Davis, R. (2005). Zipf's law, music classification and aesthetics. *Computer Music Journal*, *29*(1), 55–69.

Manaris, B., Roos, P., Machado, P., Krehbiel, D., Pellicoro, L., & Romero, J. (2007). A corpus-based hybrid approach to music analysis and composition. In *Proceedings of the 22nd conference on artificial intelligence (AAAI 07)*, Vancouver, BC.

Marchenko, Y., Chua, T.-S., & Aristarkhova, I. (2005). Analysis and retrieval of paintings using artistic color concepts. In *ICME* (pp. 1246–1249). New York: IEEE Press.

Nadal, M. (2007). *Complexity and aesthetic preference for diverse visual stimuli*. PhD thesis, Departament de Psicologia, Universitat de les Illes Balears.

Neufeld, C., Ross, B., & Ralph, W. (2007). The evolution of artistic filters. In J. Romero & P. Machado (Eds.), *The art of artificial evolution*. Berlin: Springer.

Rigau, J., Feixas, M., & Sbert, M. (2008). Informational dialogue with Van Gogh's paintings. In *Eurographics symposium on computational aesthetics in graphics, visualization and imaging* (pp. 115–122).

Romero, J., Machado, P., Santos, A., & Cardoso, A. (2003). On the development of critics in evolutionary computation artists. In R. Günther et al. (Eds.), *Lecture notes in computer science:*

Vol. 2611. Applications of evolutionary computing, EvoWorkshops 2003: EvoBIO, EvoCOMNET, EvoHOT, EvoIASP, EvoMUSART, EvoSTOC, Essex, UK. Berlin: Springer.

Saunders, R. (2001). *Curious design agents and artificial creativity—a synthetic approach to the study of creative behaviour*. PhD thesis, University of Sydney, Department of Architectural and Design Science Faculty of Architecture, Sydney, Australia.

Savarese, J. M., & Miller, R. (1979). Artistic preferences and cognitive-perceptual style. *Studies in Art Education*, *20*, 41–45.

Schmidhuber, J. (1997). Low-complexity art. *Leonardo, Journal of the International Society for the Arts, Sciences, and Technology*, *30*(2), 97–103. http://www.jstor.org/stable/1576418.

Schmidhuber, J. (1998). Facial beauty and fractal geometry. http://cogprints.org/690/.

Schmidhuber, J. (2007). Simple algorithmic principles of discovery, subjective beauty, selective attention, curiosity and creativity. In M. Hutter, R. A. Servedio & E. Takimoto (Eds.), *Lecture notes in computer science: Vol. 4754. ALT* (pp. 32–33). Berlin: Springer.

Sobel, I. (1990). An isotropic 3 × 3 image gradient operator. In *Machine vision for three-dimensional scenes* (pp. 376–379).

Spector, L., & Alpern, A. (1994). Criticism, culture, and the automatic generation of artworks. In *Proceedings of twelfth national conference on artificial intelligence* (pp. 3–8). Seattle/Washington: AAAI Press/MIT Press.

Spehar, B., Clifford, C. W. G., Newell, N., & Taylor, R. P. (2003). Universal aesthetic of fractals. *Computers and Graphics*, *27*(5), 813–820.

Staudek, T. (2002). *Exact aesthetics. Object and scene to message*. PhD thesis, Faculty of Informatics, Masaryk University of Brno.

Staudek, T. (2003). Computer-aided aesthetic evaluation of visual patterns. In *ISAMA-BRIDGES conference proceedings*, Granada, Spain (pp. 143–149).

Svangård, N., & Nordin, P. (2004). Automated aesthetic selection of evolutionary art by distance based classification of genomes and phenomes using the universal similarity metric. In R. Günther et al. (Eds.), *Lecture notes in computer science: Vol. 3005. Applications of evolutionary computing, EvoWorkshops 2004: EvoBIO, EvoCOMNET, EvoHOT, EvoIASP, EvoMUSART, EvoSTOC*, Coimbra, Portugal (pp. 445–454). Berlin: Springer.

Taylor, R. P., Micolich, A. P., & Jonas, D. (1999). Fractal analysis of Pollock's drip paintings. *Nature*, *399*, 422.

Teller, A., & Veloso, M. (1996). PADO: a new learning architecture for object recognition. In K. Ikeuchi & M. Veloso (Eds.), *Symbolic visual learning* (pp. 81–116). London: Oxford University Press. http://www.cs.cmu.edu/afs/cs/usr/astro/public/papers/PADO.ps.Z.

Tong, H., Li, M., Zhang, H., He, J., & Zhang, C. (2004). Classification of digital photos taken by photographers or home users. In K. Aizawa, Y. Nakamura & S. Satoh (Eds.), *Lecture notes in computer science: Vol. 3331. PCM (1)* (pp. 198–205). Berlin: Springer.

Tyler, C. W. (Ed.) (2002). *Human symmetry perception and its computational analysis*. Hillsdale: Erlbaum.

Uduehi, J. (1995). A cross-cultural assessment of the maitland graves design judgment test using U.S. and Nigerian students. *Visual Arts Research*, *21*(2), 11–18.

Wallraven, C., Cunningham, D. W., & Fleming, R. (2008). Perceptual and computational categories in art. In P. Brown (Ed.), *International symposium on computational aesthetics in graphics, visualization, and imaging* (pp. 131–138). Aire-la-Ville: Eurographics Association. http://computational-aesthetics.org/2008/.

Wallraven, C., Fleming, R. W., Cunningham, D. W., Rigau, J., Feixas, M., & Sbert, M. (2009). Categorizing art: comparing humans and computers. *Computers & Graphics*, *33*(4), 484–495.

Wertheimer, M. (1939). Laws of organization in perceptual forms. In W. D. Ellis (Ed.), *A source book of gestalt psychology* (pp. 71–88). New York: Harcourt Brace.

Wong, L.-K., & Low, K.-L. (2009). Saliency-enhanced image aesthetics class prediction. In *ICIP* (pp. 997–1000). New York: IEEE Press.

Yan, Y., & Jin, J. S. (2005). Indexing and retrieving oil paintings using style information. In S. Bres & R. Laurini (Eds.), *Lecture notes in computer science: Vol. 3736. VISUAL* (pp. 143–152). Berlin: Springer.

Zell, A., Mamier, G., Vogt, M., Mache, N., Hübner, R., Döring, S., Herrmann, K.-U., Soyez, T., Schmalzl, M., Sommer, T., et al. (2003). *SNNS: Stuttgart neural network simulator user manual, version 4.2* (Technical Report 3/92). University of Stuttgart, Stuttgart.

Zipf, G. K. (1949). *Human behaviour and the principle of least effort: an introduction to human ecology*. Reading: Addison-Wesley.