Finding Creativity in an Artificial Artist

David Norton, Derrall Heath and Dan Ventura

Computer Science Department

Brigham Young University

dnorton@byu.edu, dheath@byu.edu, ventura@cs.byu.edu

Abstract

Creativity is an important component of human intelligence, and imbuing artificially intelligent systems with creativity is an interesting challenge. In particular, it is difficult to quantify (or even qualify) creativity. Recently, it has been suggested that conditions for attributing creativity to a system include: appreciation, imagination, and skill. We demonstrate and describe an original computer system (called DARCI) that is designed to produce images through creative means. We present methods for evaluating DARCI and other artificially creative systems with respect to appreciation, imagination, and skill, and use these methods to show that DARCI is arguably a creative system.

Introduction

DARCI (Digital ARtist Communicating Intention) is a persistent computer system currently designed to automatically render images to match a list of adjectives. For example, if someone were to request it, DARCI could render an image of Van Gogh's Starry Night so that it could be described as happy and funny. DARCI is part of our research exploring the perception of creativity in an artificial system. In designing and evaluating DARCI we

have used Colton's creative tripod as a reference for the requirements of creativity (Colton, 2008). This paper illustrates how DARCI functions according to Colton's creative tripod and how we can evaluate the creativity of DARCI and other artificial systems.

With the creative tripod, Colton posits three behaviors required for a system to be considered creative: skill, imagination, and appreciation. These behaviors correlate strongly with the common notion of novelty and quality being central to creativity (Boden, 1999). Skill is the ability to produce quality artefacts that are recognized as members of their intended domain. Imagination is the ability to produce novel artefacts with the added caveat that they are meaningful. In other words, they are not the result of entirely random processes. Appreciation is the capability of the system to evaluate the artefacts it produces.

DARCI demonstrates appreciation by evaluating how strongly an image matches a database of adjective semantics (i.e. synsets). This is accomplished through a complex of neural networks that, given a set of features extracted from the image, output a value for each synset found in a growing database. DARCI demonstrates skill by creating images that correlate with the appropriate outputs of the neural networks. This is accomplished by applying a series of filters to the image in question. Suitable filters are discovered through a genetic algorithm. Finally, DARCI demonstrates imagination by generating unpredictable yet non-random images by using the aforementioned genetic algorithm. Details on how each of these behaviors is implemented follow.

Additionally, in order to evaluate DARCI we present a series of objective tests to evaluate appreciation and a necessarily subjective survey to evaluate the quality and novelty of DARCI's creations. We show that this survey can reliably measure what untrained human observers think of DARCI's artifacts with respect to creativity. Finally, we discuss how our results suggest that DARCI could be considered a creative system.

Background

There have been many systems designed to produce art. Some of the more prominent systems recognized in the budding field of computational creativity include Harold Cohen's

AARON (McCorduck, 1991) and Simon Colton's Painting Fool (Colton, 2011). Both of these systems are personified to some extent by their creators and some, if not all, of their inner workings are kept from the public. We see DARCI as a system analogous to these, although our intention is to disclose all of DARCI's mechanisms in our research.

Fundamental to DARCI's creation process is an evolutionary mechanism which we will describe in detail shortly. Evolutionary art, such as the images DARCI produces, has a rich background with many developments. Traditionally, evolutionary art has been an interactive affair involving the creation of a virtual space of myriad possible images for a human artist or artists to explore. This space of possible images is defined by a digital encoding, or genotype, and the rules to transform this genotype into an actual image, commonly called the phenotype. Evolutionary mechanisms begin with a population of random genotypes which are evaluated based on the qualities of their respective phenotypes. This evaluation is effected by a fitness function and determines which genotypes get to pass on their traits to future generations of the evolutionary mechanism. Since art is inherently subjective and difficult to parameterize, this evaluation is often left to human judgment. Hence a human's aesthetic sense acts as the fitness function for most evolutionary computation used for producing art (Rooke, 2002 ; Todd & Latham, 1992 ; Sims, 1991 ; Secretan et al., 2011).

Unfortunately, evolutionary mechanisms usually require large populations and many generations to converge to interesting images. This can be tiresome to human evaluators, not to mention eliminates much of the autonomy that we are interested in establishing in a creative system. Other evolutionary systems have been developed that automatically assign a fitness function to phenotypes. Most of these systems extract quantifiable features from the images to evaluate them. The hope is that aesthetic images will contain certain measureable features. Examples of features used in existing evolutionary systems include: how closely the image color distribution matches the color distribution of highly rated Flickr images (Oranchak, 2007), how complex the image is as measured by image compression (Machado & Cardoso, 2002), a geometric assessment of regions within an im-

age (Greenfield, 2002), etc.

More sophisticated approaches use a dynamic fitness function that changes based on conditions in the evolutionary environment. For example, DiPaola and Gabora have designed a system that examines two sets of features when assigning fitness to images: one is the degree to which an image is similar to a target image and the other is the degree to which an image follows quantifiable rules of aesthetics. The weight that the algorithm puts on each set of features changes whenever evolution becomes stagnant (DiPaola & Gabora, 2009). Another example of a dynamic fitness function is a co-evolutionary model implemented by Greenfield. In this algorithm, the system co-evolves a population of images and a population of image filters together. The fitness functions of each population differ, but are dependent upon the interaction between the two populations (i.e. how the image filters affect the images). Thus, the fitness functions fluctuate as if in an ecological setting (Greenfield, 2007).

Another approach to automating fitness evaluation that has been explored is to model natural systems that can be used to evaluate images. For example, one could model a human's appreciation of images and use that model as the fitness function rather than an actual human. This particular case was first explored by Baluja et al. using an artificial neural network to model the aesthetic preferences of specific users (Baluja, Pomerleau, & Jochem, 1994). To train the neural network, Beluja used data collected from a traditional human-as-the-judge evolutionary mechanism. The images that the user selected each generation were labeled positively as aesthetic images. The entire pixel space of each image was the input for the neural network. Beluja had minimal success noting the difficulty in using such an enormous input space for such a sparse selection of data points.

Recently, Machado et al. augmented their NEvAr system by incorporating a dynamic artificial neural network model to evaluate the novelty of the system's creations (Machado, Romero, & Manaris, 2007). They trained their new model to distinguish between two pools of images: those that NEvAr created and paintings by renowned artists. As suggested by Beluja, for the inputs into the neural net they used carefully selected image features rather

than the entire pixel space. Machado then used the model to act as the fitness function for NEvAr's evolutionary mechanism. Images that the system recognized as its own creations were rejected, thus the system had to evolve to change its own style. After terminating the evolutionary mechanism, the rejected creations were added to the neural network model as NEvAr's creations. The experiment was repeated again. This process was repeated twelve times forcing NEvAr to dramatically change its style.

With DARCI, we have implemented a fitness function most akin to the modeling approaches just described. As with Beluja, we use artificial neural networks to model user aesthetics; and, similar to NEvAr we use image features as input to the neural networks and dynamically update the trained model. However, DARCI is a unique system in several important ways.

Since DARCI is the subject of research in computational creativity rather than evolutionary art, DARCI is designed with a holistic approach to creativity. This means that we are interested in automating every aspect of creativity, not just the production of artifacts. The evolutionary mechanism is one piece in a bigger system that includes developing a language to express meaning to an audience, and doing so online in a social context.

In order to develop the aforementioned language, DARCI's fitness function is composed of not one but hundreds (potentially thousands) of neural networks. Each neural network corresponds to a specific adjective and is an abstract model of how humans identify that adjective in images. In other words, DARCI's fitness function is not a direct measure of one person's aesthetic sense, but a measure of an aggregate sense of what an image means.

Csíkszentmihályi explained the necessity of social context to define creativity (Csíkzentmihályi, 1996). DARCI is designed to learn and create within such a context primarily through the system's interactions with people via the internet. Ultimately, DARCI's representation of adjectives is a reflection of those who have trained the system using DARCI's website. As previously mentioned, traditional genetic algorithms require social interaction as well; however, these traditional algorithms involve the explicit creation

of artifacts as opposed to the creation of a language—in DARCI's case, a language for communicating adjectives with images.

Previous research has examined DARCI's ability to produce images that communicate specific meaning (Norton, Heath, & Ventura, 2010, 2011). While touching on that topic, this paper will focus more on DARCI's ability to produce creative images in general, as well as how we can measure that creativity.

Appreciation

Colton defines appreciation in a computational system as the ability for that system to evaluate its own artefacts. DARCI's artwork (or artefacts) comes in the form of images. In order for DARCI to appreciate art, it must first acquire some basic understanding of art. For example, in order for DARCI to appreciate an image that is dark and gloomy, DARCI must first understand the concepts *dark* and *gloomy*. To do this, DARCI must learn to associate images with artistic descriptions.

Image Feature Extraction

Before DARCI can form associations between images and descriptive words, the appropriate image features for the task must be extracted from the image. We use low-level features that can characterize the various ways that an image can be appreciated. There has been a large amount of research done in the area of image feature extraction (Gevers & Smeulders, 2000 ; Datta, Joshi, Li, & Wang, 2006 ; Li & Chen, 2009 ; Wang, Yu, & Jiang, 2006 ; King, Ng, & Sia, 2004 ; Wang & He, 2008). We select 102 image features from all of these areas of research. Our set of image features are broken down into the areas of color, light, texture, and shape as shown in Table 1.

Visuo-Linguistic Associations

DARCI forms an appreciation of art by making associations between image features and image descriptions. An image can be described and appreciated in many ways: by the subject of the image, by the aesthetic qualities of the image, by the emotions that the

Table 1: Set of features DARCI extracts from an image in order to associate images with adjectives.

Color & Light:
1. Average red, green, and blue
2. Average hue, saturation, and intensity
3. Unique hue count (20 buckets)
4. Average hue, saturation, and intensity contrast
5. Dominate hue
6. Percent of image that is the dominate hue

Shape:
1. Geometric moment
2. Eccentricity
3. Invariant moment (5x vector)
4. Legendre moment
5. Zernike moment
6. Psuedo-Zernike moment
7. Edge direction histogram (30 bins)

Texture:
1. Co-occurrence matrix (x4)
   1. Maximum probability
   2. First order element difference moment
   3. First order inverse element difference moment
   4. Entropy
   5. Uniformity
2. Edge frequency (25x vector)
3. Primitive length
   1. Short primitive emphasis
   2. Long primitive emphasis
   3. Gray-level uniformity
   4. Primitive uniformity
   5. Primitive percentage

image evokes, by associations that can be made with the image, by the meanings found within the image, and possibly others. To teach DARCI how to make associations with such descriptors, we present it with images labeled appropriately. Eventually we would like DARCI to understand images from all of these perspectives. However, for now, we have reduced descriptive labels exclusively to delineated lists of adjectives.

WordNet. We use WordNet's (Fellbaum, 1998) database of adjectives to give us a large, yet finite, set of descriptive labels. Even though our potential labels are restricted, the complete set of WordNet adjectives can allow for images to be described by their emotional effects, most of their aesthetic qualities, many of their possible associations and meanings, and even, to some extent, by their subject.

In WordNet, each word belongs to a synset of one or more words that share the same meaning. If a word has multiple meanings, then it can be found in multiple synsets. Our image classification labels actually consist of a unique synset identifier, rather than the adjectives themselves.

Data Collection. To collect training data we have created a public website for training DARCI (http://darci.cs.byu.edu). From this website, users are presented with a random image and asked to provide adjectives that describe the image. When users input a word

with multiple senses, they are presented with a list of the available senses, along with the WordNet gloss, and asked to select the most appropriate one. Additionally, for each image presented to the user, we list seven adjectives that DARCI associates with the image. The user is then allowed to flag those labels that are not accurate. This creates strictly negative examples of those synsets, which will be important in the learning process.

Learning Method. In order to make the association between image features and synsets, we use a series of artificial neural networks (ANNs) that we call the appreciation network. The appreciation network has a unique ANN, with a single output node, for each synset that has a sufficient amount of training data. For the results presented in this paper, that threshold is fifteen positive training instances. As we incrementally accumulate more data, new ANNs can be created to accommodate the new synsets. This process ensures that the synsets in question are not too obscure or accidental. It also ensures a minimum amount of training data for each synset. As of writing this paper, the appreciation network contains 211 ANNs. This means that DARCI essentially "knows" 211 synsets.

Learning image to synset associations is a *multi-label classification* problem (Tsoumakas & Katakis, 2007), meaning each image can be associated with more than one synset. We cannot assume that each training image will be labeled with all the possible correct synsets. As we train the appreciation network on an image, we only train the ANNs that are explicitly labeled (as positive or negative), and ignore the other neural nets.

ANNs require a lot of training data to converge. Currently, of the 211 synsets known to DARCI, there are on average just over 33 positive data instances per synset. In order to enhance the amount of positive and negative data used to train the appreciation network, we utilize synset relationships built into WordNet in addition to statistical correlations present in our data as described in previous work (Norton et al., 2010).

Skill and Imagination

Because Colton's notions of skill and imagination are very closely linked in our work, we will discuss them together. DARCI demonstrates skill by rendering images to match

given lists of adjectives. DARCI demonstrates imagination by creating these artefacts in a non-random way while producing unpredictable results that still reflect the original images.

Multiple definitions of skill emphasize the fact that a skill must be learned or acquired through training. Also, because of the innate ability of evolution to yield unpredictable solutions to problems, evolutionary methods are frequently used in computational creativity experiments (Gero, 1996). Thus, in order to argue for skill and imagination in DARCI's creation process, we have implemented an evolutionary mechanism to render images so that they visually express the meaning of given synsets.

Our evolutionary mechanism operates in two modes. The initial mode, which we call *practice mode*, operates by exploring the space of image filters that will render any image according to a single specific synset. For this mode, DARCI creates and maintains a separate gene pool for each synset that the system knows. The second mode, called *commission mode*, operates by exploring the space of image filters that will render a specific image according to a specified list of synsets. For this mode, users prescribe the image and list of synsets that they wish DARCI to render—in other words, they "commission" DARCI. For each commission, DARCI creates a unique gene pool that terminates once the commission is complete. For both modes, the evolutionary mechanism functions as follows.

The genotypes that comprise each gene pool are lists of filters, and their accompanying parameters, for processing an image. Many of these filters are similar to those found in Adobe Photoshop and other image editing software. Others come from a series of 1000 filters that Colton et al. discovered using their own evolutionary mechanism (Colton et al., 2010). This set of filters, called *Filter Feast*, is divided into categories of aesthetic effect that were discovered by exploring combinations of very basic filters within a tree structure. We have treated Filter Feast filters as if each category were a unique filter with a single parameter that specifies the specific filter within the category to use. Figure 1 gives an example of a genotype and its effect on a sample image. There are a total of sixty-one traditional filters that we selected for DARCI to use and a total of thirty-one categories of filters from *Filter Feast*, making ninety-two filters available for each genotype. We selected

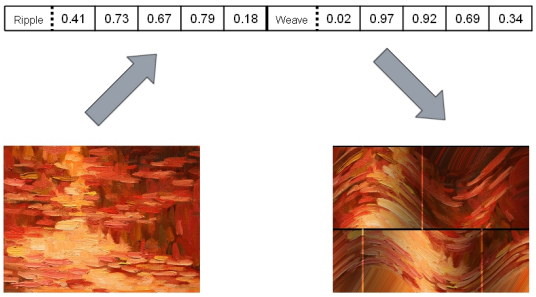


Figure 1. Sample genotype (list of image filters with parameters) and its effect on an image. "Ripple" and "Weave" are the names of two (of ninety-two) possible filters. Original image courtesy of William Meire.

traditional filters that were easily accessible, diverse, fast, and that didn't incorporate alpha values (since our feature extraction techniques cannot yet process alpha values).

The fitness function for the evolutionary mechanism can be expressed by the following equation:

$$\text{Fitness}(g) = \lambda_A A(g) + \lambda_I I(g) \tag{1}$$

where g is an image artifact and $A : G \to [0, 1]$ and $I : G \to [0, 1]$ are two metrics: appreciation and interest. These compute a real-valued score for an image artifact (here, G represents the set of all image artifacts). $\lambda_A + \lambda_I = 1$, and for now, $\lambda_A = \lambda_I = 0.5$.

Both metrics used in the fitness function are applied to the phenotype (the image that results when each genotype is applied to a source image). The fitness of every phenotype within a generation of the evolutionary mechanism is determined using the same source image; but, the source image used from generation to generation depends upon which mode the system uses. In commission mode, the source image is the same from generation to generation, while in practice mode the source image for each generation is randomly selected from DARCI's growing image database.

The appreciation metric A is computed as the (weighted sum) of the output(s) of the appropriate appreciation network(s), producing a single (normalized) value:

$$A(g) = \sum_{w \in C} \alpha_w \text{net}_w(g) \tag{2}$$

where C is the set of synsets to be portrayed, $\text{net}_w(\cdot)$ is the output of the appreciation network for synset w, $\sum_w \alpha_w = 1$, and $\alpha_w = 1/|C|$ (though this can, of course, be changed to weight synsets unequally).

The interest metric I penalizes phenotypes that are either too different from the source image, or are too similar. The metric begins by tallying the number, n, of image analysis features that have similar values between the two images (i.e. that fall within a specified distance of each other). This can be expressed with the following equation:

$$n = \sum_i \lceil 0.3 - |F_i^S - F_i^P| \rceil \tag{3}$$

F_i^S represents feature i of the source image and F_i^P represents feature i of the phenotype. Note that all features are normalized to the range [0...1], so the ceiling function above returns either 0 or 1. The value 0.3 was chosen empirically. The interest metric is calculated using n as follows:

$$I(g) = 1 - \begin{cases} \frac{\tau_d - n}{\tau_d} & \text{if } n < \tau_d \\ \frac{n - \tau_s}{|F| - \tau_s} & \text{if } n > \tau_s \\ 0 & \text{if } \tau_d \leq n \leq \tau_s \end{cases} \tag{4}$$

τ_d and τ_s are constants that correspond to the threshold for determining, respectively, when a phenotype is too different from or too similar to the source image. The values $\tau_d = 20$ and $\tau_s = 57$ were used here. $|F|$ is the total number of features analyzed, in our case 102.

Fitness-based tournament selection determines those genotypes that propagate to the next generation and those genotypes that participate in crossover. One-point "cut and splice" crossover is used to allow for variable length offspring. Crossover is accomplished in two stages: the first occurs at the filter level, so that the two genomes swap an integer number of filters; the second occurs at the parameter level, so that filters on either side of

Number of Sub-Populations	8
Size of Sub-Populations	15
Crossover Rate	0.4
Mutation Rate	0.1
Parameter Mutation Rate	0.9
Migration Rate	0.2
Migration Frequency	0.1
Tournament Selection Rate	0.75
Initial Genotype Length	2 to 4 filters

Table 2: Parameters used for the evolutionary mechanism.

the cut point swap an integer number of parameters. By necessity, parameter list length is preserved for each filter. Table 2 shows the parameter settings used.

Mutation rate is the probability that a mutation will occur in each genotype. Parameter mutation rate is the probability that when a mutation occurs, it is a parameter mutation; otherwise, it is a filter mutation. Filter mutation is a wholesale change of a single filter (discrete values), while parameter mutation is a change in parameter values for a filter (continuous values). When a parameter mutation occurs, anywhere from one to all of the parameters (uniformly chosen) for a single filter in a genotype are changed. The degree of this change, Δf_i, for each parameter, i, is determined by one of the following two equations chosen randomly with equal probability:

$$\Delta f_i = (1 - f_i) \cdot rand\left(0, \frac{(|f| + 1) - |\Delta f|}{|f|}\right) \tag{5}$$

$$\Delta f_i = -f_i \cdot rand\left(0, \frac{(|f| + 1) - |\Delta f|}{|f|}\right) \tag{6}$$

Here, $|f|$ is the total number of parameters in the mutating filter, $|\Delta f|$ is the number of changing parameters in the mutating filter, and $rand(x, y)$ is a function that uniformly selects a real value between x and y.

Because there are potentially many ideal filter configurations for modeling any given synset, we have implemented sub-populations within each gene pool. This allows the evolutionary mechanism to converge to multiple solutions, all of which could be different and

valid. The migration frequency controls the probability that a migration will occur at a given epoch, while the migration rate refers to the percentage of each sub-population that migrates. Migrating genomes are selected uniform randomly, with the exception that the most fit genotype per sub-population is not allowed to migrate. Migration destination is also selected uniform randomly, except that sub-population size balancing is enforced.

Practice gene pools are initialized with random genotypes, while commission gene pools are initialized with the most fit genotypes from the practice gene pools corresponding to the requested synsets. This allows commissions to become more efficient as DARCI practices known synsets. It also provides a mechanism for balancing permanence (artist memory) with growth (artistic progression).

Evaluation of Creativity

Because creativity is subjective, evaluation of DARCI's creative process is a difficult prospect. For the purpose of evaluation we break down creativity into two parts: the appreciation of the system in the creative process, and the creativity of the artifacts that the system produces as a function of their novelty and value. The novelty and value of the artifacts are in turn a product of the skill and imagination involved in creating them. Thus, Colton's creative tripod comes to bear in our evaluation of DARCI.

Evaluation of Appreciation

DARCI's appreciation is effectively a measure of the system's ability to label images with adjectives. While the process of determining what adjectives (or synsets) describe an image is subjective, learning these associations is essentially a multi-label classification problem that we can measure by using the following standard multi-label classification evaluation metrics (Zhang & Zhou, 2006 ; Schapire & Singer, 2000).

- *Hamming Loss* - The average percentage of correct synsets not predicted and incorrect synsets predicted.
- *Precision* - The average percentage of predicted synsets that are correct.

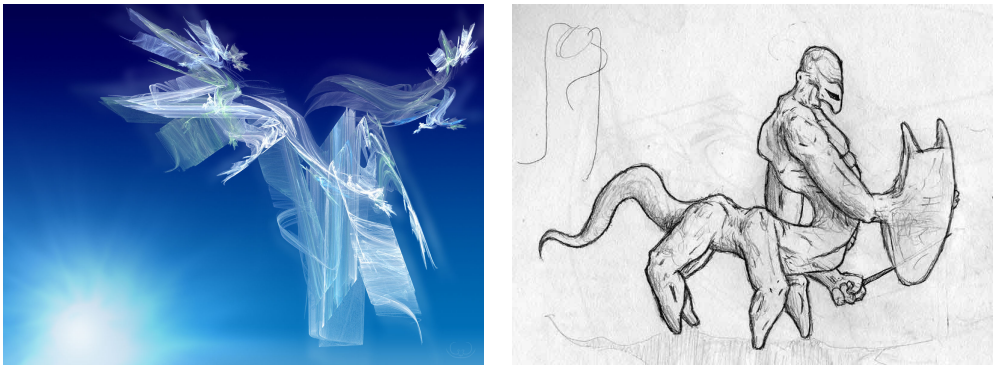
Table 3: Ten-fold cross-validation results of DARCI's appreciation network compared with a random untrained ANN. Lower numbers indicate better performance for Hamming loss, one-error, and coverage, while higher numbers are better for precision and recall.

	DARCI	Random
Hamming Loss	**0.307**	0.502
Precision	**0.606**	0.390
Recall	**0.563**	0.446
One-Error	**0.297**	0.579
Coverage	**0.347**	0.581

- *Recall* - The average percentage of correct synsets that are predicted.
- *One-Error* - The percentage of top ranked synsets that are not in the set of correct synsets.
- *Coverage* - How far, on average, we need to go down the ranked list of synsets in order to cover all the correct synsets.

Synsets are ranked according to a function of the activation values of their respective neural net output nodes (Norton et al., 2010). A synset is predicted when its corresponding output node has an activation value greater than or equal to the threshold of 0.5. We evaluate DARCI using these metrics with ten-fold cross-validation and the 2101 images currently in the DARCI database. Table 3 shows the results of the evaluation metrics comparing DARCI to a random untrained neural network.

It is important to note that these metrics are only applied to the synsets explicitly given to the images. In other words, the average image only has approximately nine synsets labeled as positive or negative (out of the 211 possible). This means we can only measure the performance on those nine synsets because we don't know anything about the remaining 202 synsets. Ideally, we would like a test set that has labels for all 211 synsets for each test image in order to fully evaluate how well DARCI can generalize. Despite this limitation, ten-fold cross-validation still gives us a good idea of how well the appreciation network is learning.



(a) blue, cold, reflective, beautiful, tranquil, glazed, patterned

(b) penciled, cold, grey, one-dimensional, icky, pastel, primitive

Figure 2. Images that DARCI has interpreted. Words underneath each image are the adjectives DARCI associated with each image. Images courtesy of Shaytu Schwandes.

Figure 2 shows DARCI describing two images on which DARCI has not been trained, and a case can be made for describing each image the way DARCI did. These results show that DARCI is making progress in being able to appreciate images. This ability to appreciate images has been shown to be important for evaluating the artefacts DARCI produces.

Evaluation of Skill and Imagination

As previously mentioned, creativity is generally considered to be a function of quality and novelty. These attributes of creativity correspond to Colton's suggested skill and imagination behaviors and are observable in the final product of creativity, the artifact itself. In order to attribute creativity to DARCI, we need to know how creative (i.e. novel and valuable) DARCI's artifacts are. Since DARCI is learning to produce images that reflect adjectives, one measure of quality could arguably be the fidelity of the adjective representation. We have studied this measure in previous work (Norton et al., 2010, 2011). While there is value to this metric, it clearly does not fully reflect creativity. First of all, there is no measure of novelty. Additionally, there is no measure of intrinsic value in the way a human critic might attribute it. In the domain of visual art there needs to be some sense of skill and aesthetics on order to attribute value to an artifact.

In order to evaluate quality more completely, while also evaluating novelty, we have devised a series of six questions for human volunteers to answer about DARCI's artifacts. These questions were designed to both explicitly and implicitly determine how an individual felt about a particular artifact from a creativity standpoint. We designed these questions as five-point Likert items (Likert, 1932); volunteers answered how strongly they agreed or disagreed (on a five point scale) with a statement as it pertained to one of DARCI's images. Following are the six statements that we used (abbreviation in parentheses):

"I like the image." (*like*)

"I think the image is novel." (*novel*)

"I would use the image as a desktop wallpaper." (*wallpaper*)

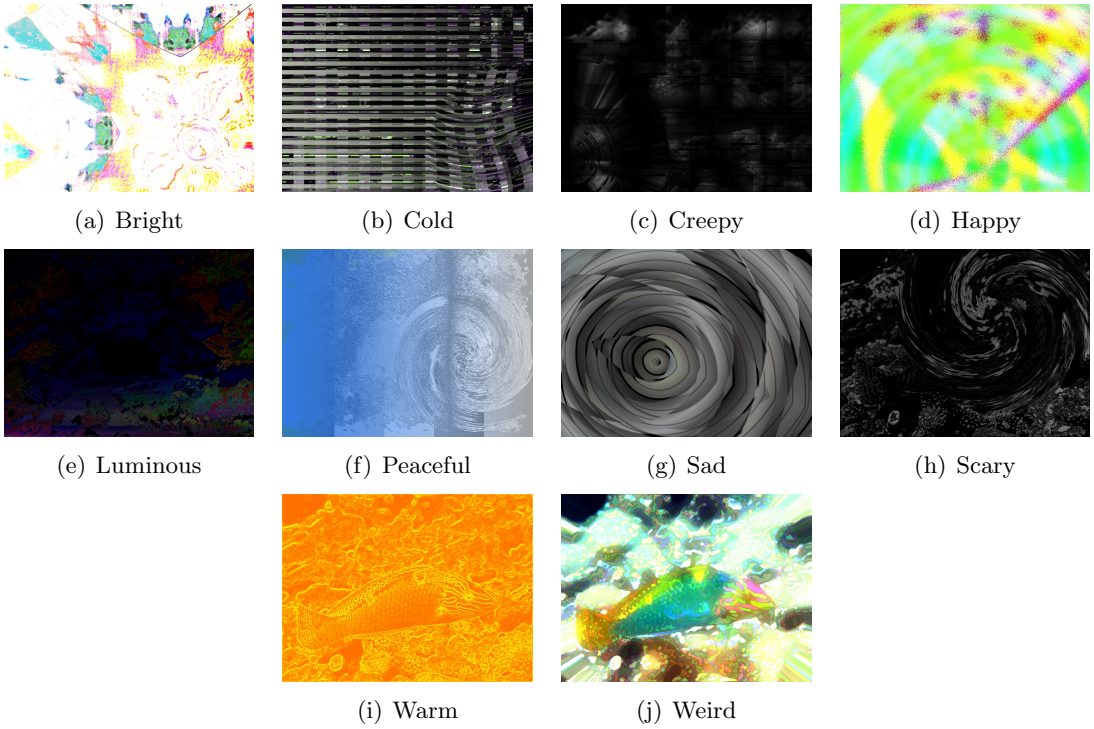
"Prior to this survey, I have never seen an image like this one." (*never seen*)

"I think the image would be difficult to create." (*difficult*)

"I think the image is creative." (*creative*)

Over the course of a week we gathered data using an anonymous online survey featuring these items. Volunteers were presented with a series of images, one at a time, and the six Likert items for each image. Each volunteer was asked to complete the Likert items for at least 10 images, though we accepted any number of images they were willing to complete in a session. Volunteers were also allowed to return to the survey for as many sessions as they wished. Due to the anonymity of the survey, multiple sessions by the same volunteer could not be matched up; however, we did ask each volunteer whether this was their first attempt at the survey or not and make use of that information in part of our analysis of the survey. There were 69 volunteers, a total of 83 sessions, and on average 7.012 images completed per session.

The images used in the online survey were randomly chosen out of 324 images created from ten commissions assigned to DARCI. Each commission consisted of the same source image, and one of ten adjectives. The adjectives we selected were: *bright*, *cold*, *creepy*, *happy*, *luminous*, *peaceful*, *sad*, *scary*, *warm*, and *weird*. These adjectives were selected



(a) Bright (b) Cold (c) Creepy (d) Happy

(e) Luminous (f) Peaceful (g) Sad (h) Scary

(i) Warm (j) Weird

Figure 3. Images that DARCI created from the same source image for ten different adjectives. These were images that received high scores from volunteers across all survey items.

because DARCI had ample training data for each and they represent a good diversity of meaning and relationships. For each commission we obtained the five best images from each of the eight sub-populations after running the commission genepools for fifty generations each. Several of the resulting 400 images were nearly or exactly identical. These were removed in order to prevent duplication artifacts in the statistical analyses we ran on the data. A high scoring example image for each of the ten adjectives is shown in Figure 3. Though the images were selected randomly, every ten images a volunteer completed in a single session consisted of an image from each of the ten unique adjectives.

For each of the ten adjectives represented, the average score for each item and the combined average score was calculated and can be found in Table 4. Figure 4 shows the average combined score for each of the adjectives in graph format for easy comparison. From this table and graph it is clear that the adjectives *weird* and *peaceful* resulted in the

	Like	Novel	Wallpaper	Never Seen	Difficult	Creative	All Items
Bright	2.509	2.246	1.912	2.316	2.316	2.702	2.333
Cold	2.697	2.909	1.939	**3.121**	2.758	3.030	2.742
Creepy	2.321	2.536	1.768	3.054	2.804	2.768	2.542
Happy	2.797	2.729	2.051	2.712	2.746	3.220	2.709
Luminous	2.596	2.404	1.842	2.579	2.439	2.719	2.430
Peaceful	**3.086**	2.931	**2.431**	2.897	**3.052**	3.259	2.943
Sad	2.821	2.643	2.286	2.536	2.554	2.804	2.607
Scary	2.966	3.085	2.169	2.932	2.881	3.271	2.884
Warm	2.473	2.364	1.582	2.691	2.218	2.527	2.309
Weird	3.068	**3.220**	2.288	3.085	3.000	**3.339**	**3.000**

Table 4: The average score of each adjective for each survey item. The item titles refer to the full statements listed in the paper. The highest scoring adjective for each item is highlighted.

most creative images according to our survey while *warm* and *bright* resulted in the least creative images.

Effectiveness of Survey

In a Likert scale survey, it is important that the items correlate with each other to some degree. Since we are interested in evaluating the creativity of DARCI's images, it is important that the six items measure creativity to some degree. In order to assess the quality of the survey items as measurements of creativity, we calculated the Cronbach alpha of the survey with respect to these items (Cronbach, 1951). Our survey received a Cronbach alpha of 0.838 indicating a high degree of consistency. In order to determine which questions were the most pertinent to the consistency of the survey, we calculated the Cronbach alpha with each question omitted. The results are shown in Table 5. From this we see that the most important item, in terms of consistency with the other items, was the statement: "I think the image is creative." Removing this statement resulted in the greatest drop in alpha value. The least important item was "Prior to this survey, I have never seen an image like this one." Removing this question actually resulted in a higher alpha than had the statement remained. In any case, every question omission still resulted in a satisfactory alpha value. From this we conclude that all of the items are valuable to our survey. In addition, since the most consistent item was the only one directly asking about creativity, we conclude that the average survey results do indeed offer a valid measure creativity.

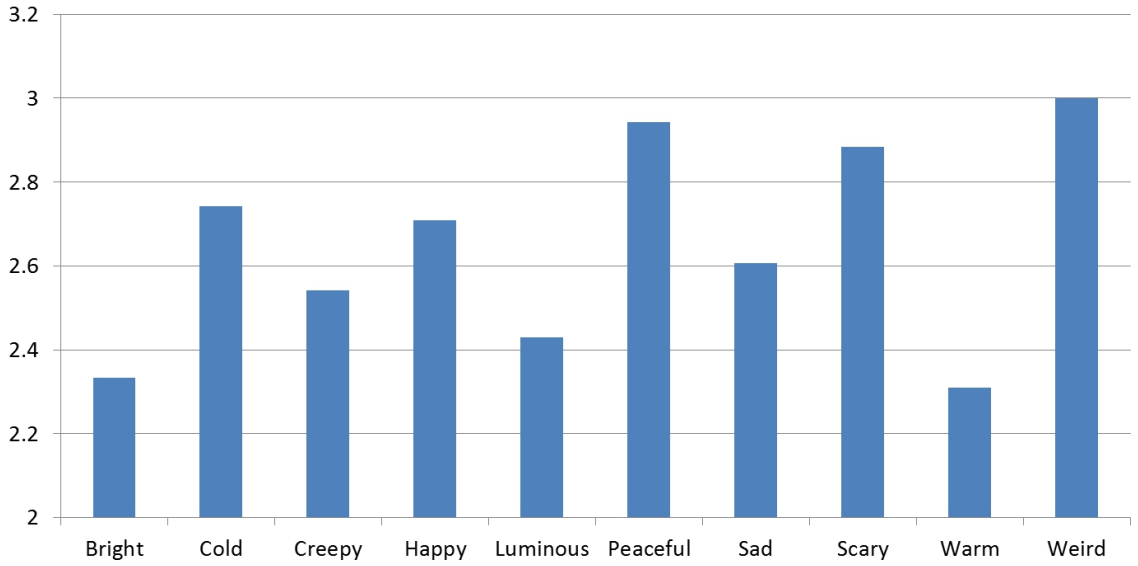


Figure 4. The average combined score across all survey items for each adjective DARCI rendered images for. This score indicates overall creativity of images produced with the indicated adjective.

The questions were valid, but how consistent were the volunteers? Our results are only useful if the volunteers were able to accurately indicate how they felt about each image. To answer this question, we calculated Cronbach's alpha with respect to each data point. Each data point corresponded to a specific volunteer and image. It was necessary to evaluate image-volunteer combinations since each image could evoke a very different response from the same person, and different volunteers could have different responses to the same image. Since there were over 300 different images, and volunteers evaluated on average just over 7

Omitted Item	Alpha Value
None Omitted	0.838
Like	0.798
Novel	0.785
Wallpaper	0.821
Never Seen	0.867
Difficult	0.808
Creative	0.782

Table 5: Alpha values measuring consistency of survey questions. The lower the alpha value, the more consistent the omitted item is with the rest of the items.

Grouping	K	Alpha Value
No Grouping	582	0.986
All Sessions	83	0.954
First Sessions	69	0.942
Adjectives	10	0.978

Table 6: Alpha values measuring consistency of volunteer responses to images. Results are shown for each grouping of the data with the resulting number of items, K.

images a piece, the chance of a volunteer evaluating the same image twice was rare and so wasn't considered.

With respect to the consistency of volunteer evaluations our survey received an alpha of 0.986. This is an extremely high score and actually raises some concerns. Cronbach's alpha is known for its susceptibility to artificial inflation due to a high number (K) of items. There were a total of 582 data points meaning that calculating the alpha with respect to these data points yields $K = 582$. This is extremely high and almost certainly has contributed to the high alpha score. In order to estimate the true alpha value without skewing due to the size of K, we grouped and then averaged data points together in various ways, and then calculated the alpha value based on these groupings.

The first grouping was by session. Each session was averaged making a total of 83 groups—still high, but much smaller than 582. This yielded an alpha of 0.954. The next grouping was by session as well but omitted all repeat sessions (sessions where the volunteer indicated that they had done the survey before). This resulted in a K of 69 and an alpha of 0.942. The final grouping was by adjective. All data points for images of the same adjective were averaged together yielding a very reasonable K of 10. The alpha for this final grouping was 0.978. While each of these groupings eliminates some important information by averaging data points, they all yield high alpha values. A summary of these results can be found in Table 6. Ultimately, these high alpha scores indicate that volunteers are indeed able to consistently articulate how they feel about each image.

Since the survey is consistent with respect to questions and volunteers, we can accept the results in Table 4 and Figure 4 as accurate indications of the creativity of DARCI's

images.

Conclusions

We have presented a system called DARCI, which is designed to produce creative artifacts in the visual arts domain, and have outlined how the system functions with respect to Colton's creative tripod (Colton, 2008). Specifically, we have illustrated how the system learns and how it creates images. We have also presented a series of widely accepted metrics that can be used to evaluate DARCI's ability to appreciate how the meanings of adjectives are expressed in images; in other words, the degree to which DARCI can identify the adjectives that describe images. Finally, we have described a survey that attempts to measure the novelty and quality of DARCI's artifacts.

The results of the appreciation metrics tell us that DARCI is indeed learning how to associate meaning to image features. The results of the online survey tell us that individuals do consider at least some of DARCI's creations to be creative. A more detailed analysis of the survey results tell us other interesting things about DARCI's creative process. We see that certain adjectives are more conducive to generating creative images than others. Five of the top seven adjectives that scored highest on average across all the questions are adjectives that describe emotion (peaceful, scary, happy, sad, and creepy). While the adjectives that scored the lowest are much simpler adjectives that only describe specific attributes (bright, warm, luminous). For example, DARCI's images that were generated to convey "peaceful" were considered more creative than those that were generated to convey "luminous". This makes sense intuitively. It is expected that images created to convey more meaningful words would be considered more novel and valuable than images created to convey simpler words. These results verify that DARCI is capable of conveying some level of meaning in the images that it creates.

It is interesting to note that the highest scoring adjective is "weird". While not usually considered an emotion, it is the only adjective of our ten with a meaning that has some obvious overlap with certain components of creativity, such as novelty. This again provides

evidence that DARCI is capable of learning words and can intentionally create artifacts that communicate those words. As a corollary, these results also support the notion that communicating intention is an important component of creativity. Artifacts that convey the most meaning are generally considered more creative than artifacts that convey little meaning.

We have shown that our image survey can help us with the subjective task of evaluating creativity. We have also shown that the survey questions are both consistent and relevant to the creative evaluation of an artifact, whether they are implicitly or explicitly dealing with creativity. Future work will involve using this survey to compared DARCI's artifacts with those created by human artists. Future work will also involve modifying and extending DARCI's capabilities to more accurately model creativity on a computer.

Références

Baluja, S., Pomerleau, D., & Jochem, T. (1994). Towards automated artificial evolution for computer-generated images. *Connection Science*, *6*, 325-354.

Boden, M. A. (1999). Handbook of creativity. In (chap. 18). Press Syndicate of the University of Cambridge.

Colton, S. (2008). Creativity versus the perception of creativity in computational systems. *Creative Intelligent Systems: Papers from the AAAI Spring Symposium*, 14-20.

Colton, S. (2011). The painting fool: Stories from building an automated painter. In J. McCormack & M. d'Inverno (Eds.), *Computers and Creativity.* Springer-Verlag.

Colton, S., Gow, J., Torres, P., & Cairns, P. (2010). Experiments in objet trouvé browsing. *Proceedings of the 1st International Conference on Computational Creativity*, 238-247.

Cronbach, L. (1951, September). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334.

Csíkzentmihályi, M. (1996). *Creativity.* HarperCollins Publishers.

Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2006). Studying aesthetics in photographic images using a computational approach. *Lecture Notes in Computer Science*, *3953*, 288-301.

DiPaola, S., & Gabora, L. (2009). Incorporating characteristics of human creativity into an evolutionary art algorithm. *Genetic Programming and Evolvable Machines*, *10*(2), 97-110.

Fellbaum, C. (Ed.). (1998). *Wordnet: An electronic lexical database*. The MIT Press.

Gero, J. S. (1996). Creativity, emergence, and evolution in design. *Knowledge-Based Systems*, *9*, 435-448.

Gevers, T., & Smeulders, A. (2000). Combining color and shape invariant features for image retrieval. *IEEE Transactions on Image Processing*, *9*, 102-119.

Greenfield, G. (2002). Color dependent computational aesthetics for evolving expressions. In R. Sarhangi (Ed.), *Bridges: Mathematical Connections in Art, Music, and Science; Conference Proceedings* (p. 9-16). Winfield, KS: Central Plains Book Manufacturing.

Greenfield, G. (2007). Co-evolutionary methods in evolutionary art. In J. Romero & P. Machado (Eds.), *The Art of Artificial Evolution: A Handbook on Evolutionary Art and Music* (p. 357-380). Berlin: Springer.

King, I., Ng, C. H., & Sia, K. C. (2004). Distributed content-based visual information retrieval system on peer-to-pear network. *ACM Transactions on Information Systems*, *22*(3), 477-501.

Li, C., & Chen, T. (2009). Aesthetic visual quality assessment of paintings. *IEEE Journal of Selected Topics in Signal Processing*, *3*, 236-252.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22*(140), 1-55.

Machado, P., & Cardoso, A. (2002). All the truth about NEvAr. *Applied Intelligence, Special Issue on Creative Systems*, *16*, 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* (p. 381-415). Berlin: Springer.

McCorduck, P. (1991). *Aaron's code: Meta-art, artificial intelligence, and the work of harold cohen*. W. H. Freeman & Co.

Norton, D., Heath, D., & Ventura, D. (2010). Establishing appreciation in a creative system. *Proceedings of the 1st International Conference on Computational Creativity*, 26-35.

Norton, D., Heath, D., & Ventura, D. (2011). Autonomously creating quality images. *Proceedings of the 2nd International Conference on Computational Creativity*, 10-15.

Oranchak, D. (2007). Evolutionary synthesis of photographic artwork using human fitness function derived from web-based social networks. *Proceedings of the 9th Annual Conference on Genetic*

and Evolutionary Computation, 2264.

Rooke, S. (2002). Eons of genetically evolved algorithmic images. In P. J. Bentley & D. W. Corne (Eds.), *Creative Evolutionary Systems* (p. 339-365). Morgan Kaufmann Publishers.

Schapire, R. E., & Singer, Y. (2000). Boostexter: A boosting-based system for text categorization. *Machine Learning*, *39*(2/3), 135–168.

Secretan, J., Beato, N., DAmbrosio, D. B., Rodriguez, A., Campbell, A., Folsom-Kovarik, J. T., et al. (2011). Picbreeder: A case study in collaborative evolutionary exploration of design space. *Evolutionary Computation Journal*, to appear.

Sims, K. (1991). Artificial evolution for computer graphics. *Computer Graphics*, *25*(4), 325-327.

Todd, S. J. P., & Latham, W. (1992). *Evolutionary art and computers.* Academic Press.

Tsoumakas, G., & Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, *3*(3), 1-13.

Wang, W.-N., & He, Q. (2008). A survey on emotional semantic image retrieval. *Proceedings of the International Conference on Image Processing*, 117-120.

Wang, W.-N., Yu, Y.-L., & Jiang, S.-M. (2006). Image retrieval by emotional semantics: A study of emotional space and feature extraction. *IEEE International Conference on Systems, Man, and Cybernetics*, *4*, 3534-3539.

Zhang, M.-L., & Zhou, Z.-H. (2006). Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering*, *18*(10), 1338-1351.