Autonomously Managing Competing Objectives to Improve the Creation and Curation of Artifacts

David Norton, Derrall Heath, Dan Ventura
Computer Science Department
Brigham Young University
Provo, UT 84602 USA
dnorton@byu.edu, dheath@byu.edu, ventura@cs.byu.edu

Abstract

DARCI (Digital ARtist Communicating Intention) is a creative system that we are developing to explore the bounds of computational creativity within the domain of visual art. As with many creative systems, as we increase the autonomy of DARCI, the quality of the artifacts it creates and then curates decreases—a phenomenon Colton and Wiggins have termed the latent heat effect. We present two new metrics that DARCI uses to evolve and curate renderings of images that convey target adjectives without completely obfuscating the original image. We show how we balance the two metrics and then explore various ways of combining them to autonomously yield images that arguably succeed at this task.

Introduction

There has been a recent push in computational creativity towards fully autonomous systems that are perceived as creative in their own right. One of the most significant problems facing modern creative systems is the level of curation that is occurring in these systems. If a system is producing dozens, hundreds, or even thousands of artifacts from which a human is choosing a single valued artifact, then is the system truly fully autonomous? Colton has argued that for a system to be perceived as creative, it must demonstrate appreciation for its own work (Colton 2008). A strong implication of this is that the system must be able to do its own curation by autonomously selecting an artifact for human judgment.

DARCI (Digital ARtist Communicating Intention) is a creative system that we are developing to explore the bounds of computational creativity within the domain of visual art. DARCI is composed of several subsystems, each with its own creative potential, and each designed to perform an integral step of image creation from conception of an idea, to design, to various phases of implementation, to curation. The most complete subsystem, and the one that is the focus of this paper, is called the *image renderer*. The image renderer uses a genetic algorithm to discover a sequence of image filters that will render an image composition (produced by another subsystem) so that it will reflect a list of adjectives (selected from yet another subsystem). After evolving a population of candidate renderings, the image renderer must select an interesting candidate that reflects both the original image *and* the given adjectives—in other words, it must curate the finished artifacts.

Historically, DARCI has been successful at producing such images when curation is a joint effort between DARCI and a human (Norton, Heath, and Ventura 2011b; Heath, Norton, and Ventura 2013). In these cases, DARCI selects a number of artifacts, and a human chooses their favorite from that selection. When DARCI curates on its own, the results have been significantly less successful. This decrease in quality is to be expected and is a phenomenon Colton and Wiggins call the *latent heat effect*—"as the creative responsibility given to a system increases, the value of its output does not (*initially*) increase ..." (emphasis added) (Colton and Wiggins 2012). Since we know DARCI is capable of producing interesting images, we are interested in *increasing* the value of the artifacts the system produces when curating alone, thus decreasing the latent heat effect.

DARCI's image renderer uses a combination of two conflicting metrics as a fitness function to evaluate and assign fitness scores to candidate artifacts. The fitness score not only drives the evolution of artifacts using a genetic algorithm, it is also used to curate the population of candidate artifacts when evolution is complete. For this paper we have made improvements to the fitness function in order to improve the quality of artifacts DARCI produces.

Previously, the fitness function has been the combined average of an ad-hoc interest metric and an adjective matching metric. In this paper, we will abandon the interest metric in favor of a new similarity metric, and combine it with an improved adjective matching metric. While we take measures to ensure that both metrics output real values in a similar range, experience has shown that the two metrics are not measuring attributes of equal quality. This has led to the observation that if combining metrics with an average, the algorithm will give disproportionate weight to the metric that is easier to maximize. Thus, we will investigate different means of combining these two metrics in an attempt to more effectively balance the requirements put upon the image rendering subsystem and decrease the latent heat effect. We show the results of these new fitness functions in figures curated strictly by DARCI.

Image Rendering

The image rendering subsystem uses a series of image filters to render pre-existing images which we refer to as *source images*. The subsystem has access to Photoshop-like filters

with varying parameters. It uses a genetic algorithm to discover the configuration and parameter settings of these image filters so that candidate artifacts will reflect target adjectives without over or under-filtering the source image (Norton, Heath, and Ventura 2011b; 2013). A genetic algorithm is used because evolutionary approaches elegantly facilitate the creation of artifacts through both combination and exploration, two processes described by Boden for generating creative products (Boden 2004). Gero has also outlined how the processes underlying evolution are ideal for producing novel and unexpected solutions, a crucial part of creativity (Gero 1996). Finally, we have shown how evolutionary algorithms approximate some aspects of the creative process in human artists (Norton, Heath, and Ventura 2011a).

In this section we will describe in detail the two metrics used in this paper: adjective matching and similarity.

Adjective Matching

The adjective matching metric is the output of a learning subsystem of DARCI called the Visuo-Linguistic Associator (VLA). The VLA is a collection of artificial neural networks (ANN) that learns to associate image features with adjectives through backpropagation training. The original VLA has been described in detail previously (Norton, Heath, and Ventura 2010). Here we introduce an improved VLA.

While DARCI is designed to function as an online system, the original VLA required subsystem resets whenever it was time to introduce new training data, essentially learning in batch. Thus, in order for DARCI to adapt, human intervention was needed at regular intervals. The new VLA uses an approach closer to incremental learning to better facilitate the desired autonomous online functionality. Additionally, the new VLA uses a more accurate and complete approach to predicting additional training data. In this section we will describe the new VLA without any assumptions that the reader is familiar with the previous system.

Training Data Training data for DARCI is contained in a database. Each data point consists of an adjective (the label), the sentiment toward the adjective (positive or negative), the image features associated with the adjective (the image), and a time stamp. In our research, the term *adjective* always refers to a unique adjective synset as defined in WordNet (Fellbaum 1998). Hence, different senses of the same word will belong to different synsets, or adjectives.

Data points are added to the database as they are submitted by volunteers using a training website (Heath and Norton 2009). Whenever the training algorithm is invoked, new *relevant* data points are introduced to the learner one at a time in the submitted order. The learner consists of a series of binary ANNs, one for each *relevant* adjective. An adjective, and any corresponding data point, is considered *relevant* once there are at least ten *distinct* positive and ten *distinct* negative instances of the adjective in the database. Here, *distinct* means occurrences of the adjective with unique sets of image features (i.e. if an adjective is used to label the same image multiple times it only counts as one occurrence). At the moment the learner is invoked, a new neural network is created for any new adjectives that have become relevant.

Table 1: Image features used to train neural networks.

Color & Light:
1. Average red, green, and blue
2. Average hue, saturation, and intensity
3. Saturation and intensity contrast
4. Unique hue count (from 20 quantized hues)
5. Hue contrast
6. Dominant hue
7. Dominant hue image percent

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

The reason we only create and train the learner on relevant data points is a matter of practicality. There are over 18000 adjective synsets in WordNet, and at the time of this writing more than 6000 adjective synsets in DARCI's database. However, most of the adjectives in DARCI's database are rare with only one or two positive data points. This is not enough data to successfully train any learner in a complex domain such as image annotation. Since performance speed is important for DARCI, accessing 6000 neural nets, most of which would be insufficiently trained, to annotate an image is impractical. As of this writing, DARCI has 237 relevant adjectives, a much more useful and manageable number. Taking synonyms into consideration, these relevant adjectives cover most standard adjectives.

The learner's neural networks are trained using standard back propagation with 102 image features as inputs. These image features are widely accepted global features for content based image retrieval, and most of them are available through the DISCOVIR (DIStributed COntent-based Visual Information Retrieval) system (King, Ng, and Sia 2004; Gevers and Smeulders 2000). A summary of the features we use can be found in Table 1. These features describe the color content, lighting, textures, and shape patterns found in images. Specific to the art domain, several researchers have shown that such features are useful in classifying images according to aesthetics (Datta et al. 2006), painting genre (Zujovic, Gandy, and Friedman 2007), and emotional semantics (Wang, Yu, and Jiang 2006). As many of these researchers have found color to be particularly useful in classifying images, we added four color-based features inspired by Li's own colorfulness features (Li and Chen 2009) to those contained in DISCOVIR. In Table 1 these colorfulness features are "Color & Light" numbers 4-7.

When training neural networks in batch, back propagation requires many epochs of training to converge. During each epoch, all of the training data is presented to the neural network in a random order. To imitate this with incremental learning, each new data point is introduced to the appropriate neural network along with a selection of previous data points. Along with this *recycled* data, additional data points

are *predicted* from the co-occurrences of adjectives with images. By including predicted data we are able to augment the limited data we do have. Similar, but less complete, approaches to augmenting training data have been successful in the past (Norton, Heath, and Ventura 2010).

Recycling Data For each new data point presented to a neural network for a given adjective, a, n positive data points from the set of all previous positive data points for the given adjective, D_{a+}, and n negative data points from the set of all previous negative data points for the given adjective, D_{a-}, are selected. The data points are selected with replacement according to the probability $P(rank(d))$ where $d \in D_{as}$, s is the sentiment of the set ($-$ or $+$), and $rank(d)$ is the temporal ordering of element d in D_{as}. The most recent element has a rank of $|D_{as}|$ and the oldest element has a rank of 1. The equation for $P(rank(d))$ is as follows:

$$P(rank(d)) = \frac{rank(d)}{\sum_{i=0}^{|D_{as}|} i} \quad (1)$$

The value for the number of previous data points chosen, n, is defined by $n = min(r, |D_{a+}|, |D_{a-}|)$ where r is a parameter setting the maximum number of data points to recycle each time a new data point is introduced. For the experiments in this paper, this value is set to 100.

Informally, every time a new data point is presented to a neural network, an equal number of positive and negative data points are selected from the previous data points for that neural network. These are selected randomly but with a higher probability given to more recent data.

Predicting Data To augment the training data we collect from DARCI's website, we analyze the co-occurrence of relevant adjectives to predict additional data points. Here we say that two adjectives co-occur whenever the same image is labeled with both adjectives at least once—these labels can be negative or positive. As each new data point is introduced to the learner, co-occurrence counts (distinct images) are updated for all pairings of relevant adjectives across all four combinations of sentiment. For example, as of this paper, 'scary' has 26 co-occurrences with 'disturbing' (or 'scary' co-occurs with 'disturbing' in 26 distinct images) and 0 co-occurrences with 'not disturbing', while 'not scary' has 5 co-occurrences with 'disturbing' and 32 co-occurrences with 'not disturbing'.

Once the co-occurrence counts have been updated, they are used to predict m positive and m negative data points to augment the new data point. m is calculated as $\lfloor pn \rfloor$ where p is a prediction coefficient and n is defined above. For this paper, p is set to 0.3. These predicted data points are not added to the database.

To predict new data points for the given adjective, a, the system first calculates each of the likelihoods that an image will be labeled with a or $\neg a$ given that the image is labeled positively or negatively with each of the adjectives, a_i, in A, the set of all relevant adjectives. Likelihood is calculated as:

$$L(a|a_i) = \frac{co(a, a_i)}{supp(a_i)} \quad (2)$$

where $co(a, a_i)$ is the co-occurrence count for a and a_i, and $supp(a_i)$ is the support of a_i (i.e. number of distinct images labeled with a_i).

Predicted data points for a are chosen using two *probability distributions* created from the above likelihoods, one for positive data points and the other for negative. The positive probability distribution is created by choosing the set of likelihoods, Λ_+, that is the set of all likelihoods described with $L(a|a_i)$ and $L(a|\neg a_i)$ that are greater than some threshold, γ, and less than 1. In this paper, γ is set to 0.4. A likelihood of 1 is omitted because it is guaranteed that there will be no new images to predict with label a. The positive probability distribution is then created by normalizing Λ_+. The negative probability distribution is created in the same way except using the set of all likelihoods, Λ_-, described with $L(\neg a|a_i)$ and $L(\neg a|\neg a_i)$ satisfying the same conditions.

For each data point to be predicted, a likelihood distribution from either Λ_+ or Λ_- is selected using the above probability distributions. Then an image is selected, using a uniform distribution, from all those images with the likelihood's label (either a_i or $\neg a_i$) that are not labeled with a. The label for the new predicted data point is a, the sentiment is the sentiment of the distribution Λ, and the features are the image features of the selected image.

Informally, data points are predicted by assuming that images labeled with adjectives that frequently co-occur with a given adjective, can also be labeled with the given adjective.

Artificial Neural Networks Once recycled and predicted data points for a particular incoming data point are selected, they are shuffled with the incoming data point and given as inputs into the appropriate neural network. The incoming data point then immediately becomes available as historical data for subsequent training data. This process is repeated for each new data point introduced to the learner. Assuming that there is sufficient data, each new data point will be accompanied by a total of $2n + 2m$ data points. In the case of this paper, that's 260 recycled or predicted data points evenly balanced between positive and negative sentiments.

As previously mentioned, one binary artificial neural network is created for each relevant adjective. These neural networks have 102 input nodes for the image features previously described. For this research, based on preliminary experimentation, the neural networks have 10 hidden nodes, a learning rate of 0.01, and a momentum of 0.1.

When the VLA is accessed for the adjective matching metric, the candidate artifact being evaluated is analyzed by extracting the 102 image features. These features are then presented to the appropriate neural network and the output is used as the actual metric. Thus, as Baluja and Machado et al. have done previously, we essentially build and use a model of human appreciation to guide the creation process so that we will hopefully produce images that humans can value (Baluja, Pomerleau, and Jochem 1994; Machado, Romero, and Manaris 2007). Unlike Baluja and Machado however, our model associates images with language and meaning (adjectives), an important step in building a system that communicates intention with its artifacts.

Similarity

The similarity metric borrows from the growing research on *bag-of-visual-word* models (Csurka et al. 2004; Sivic et al. 2005) to analyze local features rather than global ones as we have done previously (Norton, Heath, and Ventura 2011b). Typically, these local features are descriptions of points in an image that are the most surprising, or said another way, the least predictable. After such an interest point is identified, it is described with a vector of features obtained by analyzing the region surrounding the point. *Visual words* are quantized local features. A dictionary of visual words is defined for a domain by extracting local interest points from a large number of representative images and then clustering them (typically with k-means) by their features into k clusters, where k is the desired dictionary size. With this dictionary, visual words can be extracted from any image by determining to which clusters the image's local interest points belong. A bag-of-visual-words for the image can then be created by organizing the visual word counts for the image into a fixed vector. This model is analogous to the bag-of-words construct for text documents in natural language processing. These fixed vectors can then be compared to determine image similarity.

For the similarity metric used in this paper, we use the standard SURF (Speeded-Up Robust Features) detector and descriptor to extract interest points and their features from images (Bay et al. 2008). SURF quickly identifies interest points using an approximation of the difference of Gaussians function, which will often identify corners and distinct edges within images. To describe each interest point, SURF first assigns an orientation to the interest point based on surrounding gradients. Then, relative to this orientation, SURF creates a 64 element feature vector by summing both the values and magnitudes of Haar wavelet responses in the horizontal and vertical directions for each square of a four by four grid centered on the point.

We build our visual word dictionary by extracting these SURF features from more than 2000 images taken from the database of images we've collected to train DARCI. The resulting interest points are then clustered into a dictionary of 1000 visual words using Elkan k-means (Elkan 2003).

Similarity is determined by comparing candidate artifacts with the source image. We create a normalized bag-of-visual-words for the source image and each candidate artifact using our dictionary, and then calculate the *angular similarity* between these two vectors. Angular similarity between two vectors, A and B, is calculated as follows:

$$similarity = 1 - \frac{cos^{-1}(\frac{A \cdot B}{\|A\|\|B\|})}{\pi} \quad (3)$$

This metric effectively measures the number of interest points that coincide between the two images by comparing the angle between vectors A and B. In text analysis, *cosine similarity* (the parenthetical expression contained in Equation 3) is typically used to compare the similarity of documents. With this metric, as the sparseness of vectors increases, the similarity between arbitrary vectors approaches 0. In our case, as vectors are quite sparse, artifacts that are even slightly different from the source would have low scores using this measure. Nevertheless, creating renderings that are very similar to the source image is trivial as it requires simply using fewer and less severe filters. Thus, despite encountering low scores from only small differences, the genetic algorithm would be able to easily converge to near perfect or even perfect scores. This interplay between a harsh similarity metric and relative ease of convergence would place too much weight on the similarity metric. In fact, auxiliary experiments have shown that when using cosine similarity, the adjective matching metric is almost ignored in artifact production.

Since the bag-of-visual-word vectors can only contain positive values, using angular similarity instead of cosine similarity naturally constrains the output to between 0.5 and 1.0. This smaller spread in potential scores significantly reduces the negative impact of sudden jumps in similarity score due to small changes in the candidate renderings. It should be noted that in cases where a candidate artifact has no detected interest features ($\|B\| = 0$), the similarity will default to 0. This is the only case where the similarity score can be below 0.5 as the metric cannot make a comparison.

Experimental Design

Six fitness functions are explored in this paper. They are referred to as *similarity*, *adjective*, *average*, *minimum*, *alternate*, and *converge*. *Similarity* and *adjective* are the similarity and adjective matching metrics in isolation. The other four combine these two conflicting metrics in different ways. *Average* is the approach we have used in the past. With this approach, the two metrics are averaged together with equal weight. With *minimum*, the fitness function is the minimum of the metrics. *Alternate* uses one metric at a time for the fitness function, but it alternates between the two every generation beginning with adjective matching. Finally, *converge* also uses one metric at a time; however, it alternates every 20 generations also beginning with adjective matching.

The two conflicting metrics result in a process that is arguably transformational in nature, at least to a limited degree. Boden describes transformational creativity as that which transforms the conceptual space of a domain (Boden 1999). While the space of possible artifacts cannot change (the filters available for rendering images do not change), the evaluation of the artifacts does change through the interplay of the two metrics. This interplay occurs organically in the *minimum* fitness function by forcing the system to emphasize the metric that it is struggling to optimize at any given epoch during the evolutionary algorithm. The interplay of divergent metrics occurs more mechanically in the *alternate* and *converge* fitness functions by scheduling the emphasis; however, the sudden shift in metric could result in more *unexpected* results, a criterion of creativity emphasized by Maher (Maher 2010; Maher, Brady, and Fisher 2013). The scheduled approaches were inspired by Dipaola and Gabora's work with "Evolving Darwin's Gaze", an installation that also evolves images under two shifting criteria (DiPaola and Gabora 2009). Their criteria are a pixel matching metric comparing artifacts to a specific portrait of Charles Darwin, and an artistic heuristic. We anticipate that our less

Figure 1: The three source images used in all experiments. Images A and C have resolutions of 1600x1200. Image B has a resolution of 1920x1200.

restrictive metrics will ultimately allow for even more surprise and variation in artifacts, while also communicating meaning (adjectives).

Each of the above fitness functions except for *similarity* was run on three source images across five adjectives for a total of fifteen experiments per approach. *Similarity* was only run once for each source image since no adjective was needed. For algorithmic efficiency, the artifacts produced in the experiments were scaled down to a maximum width of 800 pixels. Each experiment ran for 100 generations.

The five adjectives used were 'happy', 'sad', 'fiery', 'wet', and 'peaceful'. These were chosen because they were well represented in our adjective matching training data and because they depict a range of distinct meanings and emotional valence. The three source images (referred to as images A, B, and C) are shown in Figure 1 with their corresponding resolutions.

As mentioned previously, optimizing to the similarity metric alone is trivial for the genetic algorithm since it need only remove filters to do so. However, there is no such trivial approach to optimize to the adjective metric. Historically, near perfect similarity scores are common, while near perfect adjective matching scores are non-existent. In order to balance the quality of the two metrics in our experiments, the source images were not scaled down to match the resolution of the artifacts. A source image and its otherwise unaltered counter part will yield similar but not identical visual-bags-of-words when analyzed for the similarity metric. This means that the genetic algorithm will no longer be able to trivially achieve perfect similarity. The similarity scores of each source image compared to the scaled down version of itself are, for images A, B, and C respectively: 0.826, 0.739, and 0.843 with an average score of 0.803. This means that for our experiments, the range of similarity is now more or less between 0.5 and 0.803—with a now soft ceiling. This is much closer to the range we have seen from adjective matching in auxiliary experiments: 0.144 to 0.714.

Results

In this section we will discuss DARCI's artifact selection for each experiment. While all interpretations of the images themselves are clearly subjective, we attempt to be conservative and consistent in our observations. We will discuss the artifacts in terms of the objectives of the image rendering subsystem: to depict the source image and adjective together in an *interesting* way. By *interesting* we specifically

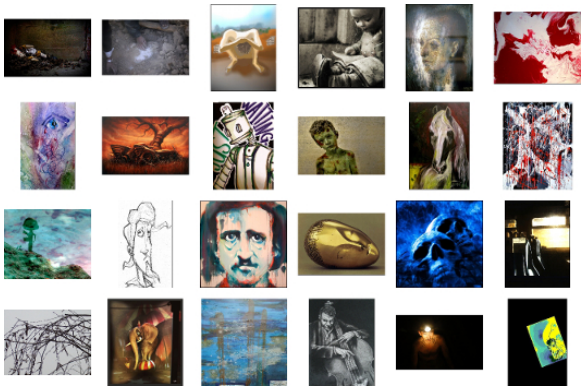

Figure 2: Sample 'sad' images from training data.

mean that extensive filtering (more than basic color filtering or use of inconspicuous filters) has occurred without removing all trace of the source image. Any hint of the source image will be considered acceptable in attributing *interest* to an artifact.

This definition of *interesting* is derived from two commonly proposed requirements for creativity applied to the specific goal of DARCI's image rendering subsystem. These two requirements are, as defined by the American Psychological Association, functionality and originality; or, as Boden described them for the domain of computation, quality and novelty (Boden 1999). Since the purpose of the image renderer is to alter a source image, elimination of the source image would not be functional. Ritchie describes a related requirement that is also applicable here—that of typicality (Ritchie 2007). Ritchie defines typicality as the extent to which an artifact is an example of its intended class. In our case this would be a rendering of a source image as opposed to an entirely new image. The second requirement, novelty, requires that the image renderer produce renderings that are distinctive. Thus, minor or no changes to a source image would clearly suggest a failure at novelty. In an attempt to reduce the amount of subjectivity in our analysis, DARCI's artifacts are either *interesting* by this definition or not. There is no attempt to rate the degree of *interest*.

In addition to being *interesting*, DARCI's artifacts must match the intended adjective. In order to be as objective as possible, we will compare DARCI's artifacts to images from the VLA training data for each given adjective. These images are representative of the types of images one would find if searching google images for a specific adjective. Examples of these images can be found in Figures 2-6. Since DARCI is rendering, as opposed to composing, and due to the limitations of DARCI's image analysis features (and indeed the limitations of the entire field of computer vision), we will be looking for similarities in color, light, and texture as opposed to similar object content.

The 'sad' training images (Figure 2) tend to be desaturated, even black and white, and/or dark with an emphasis on dull colors. The 'happy' training images (Figure 3) trend towards bright and colorful, often containing a full spectrum of colors. The 'fiery' training images (Figure 4)

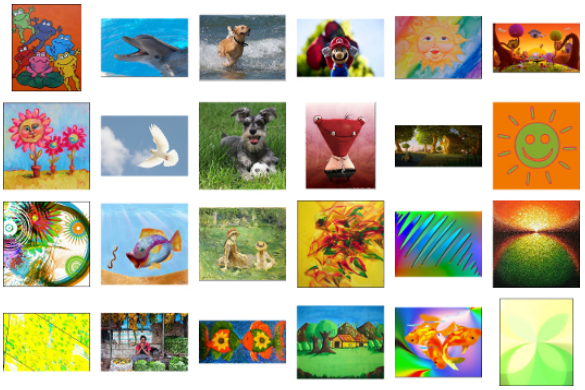

Figure 3: Sample 'happy' images from training data.

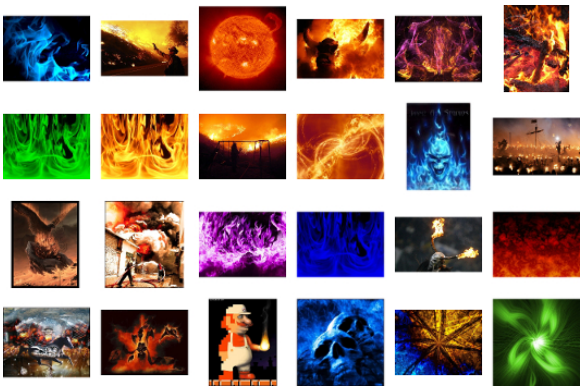

Figure 4: Sample 'fiery' images from training data.

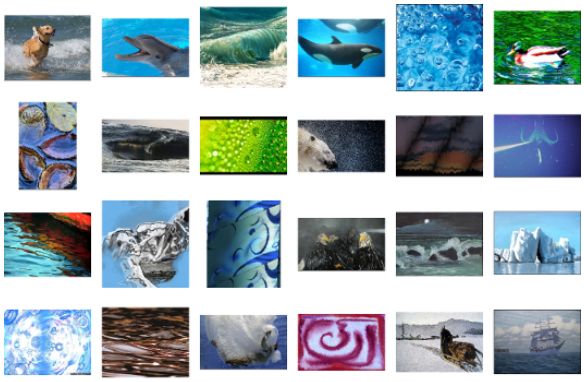

Figure 5: Sample 'wet' images from training data.

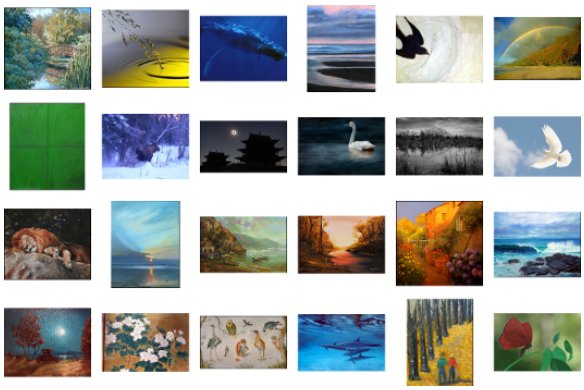

Figure 6: Sample 'peaceful' images from training data.

usually have distinct flame textures, are bright, and most are monochromatic—typically orange. The 'wet' training images (Figure 5) consist of cool colors, usually blue, and have frequent specular highlights and/or wavy patterns. Finally, the 'peaceful' training images (Figure 6) contain a variety of soft or pastel colors with a lot of smooth textures.

Ideally, the most fit artifact discovered by the genetic algorithm should be the one that best satisfies the objectives for object rendering outlined above. Thus, for most of the fitness functions, we used this method of selection. However, we anticipated that for two of the fitness functions, *alternate* and *converge* this would not be an appropriate approach. The reason for this is that both of these fitness functions only use one metric at a time, meaning that the most fit artifact discovered could only have been optimized for a single metric. The expected result would be the same as a selection from one of the control fitness functions—not an ideal balance of metrics.

We will first discuss the results of the fitness functions that use the most-fit selection process: *similarity*, *adjective*, *average*, and *minimum*. Later we will discuss *alternate* and *converge* using a different selection criteria. We will evaluate each selection process by the proportion of artifacts that meet the *interest* and adjective matching requirements.

Most Fit Selection

The most fit artifact discovered for each source image in the *similarity* control experiments is shown in Figure 7. The most fit artifact discovered in each of the other experiments is shown in Figures 8-12.

First looking at the *similarity* results (Figure 7), we see that with the exception of image A, DARCI did not select nearly identical images as we might have expected. This illustrates the effect of not scaling the source images. The chosen artifacts actually had slightly higher fitness scores than the strictly scaled down source images demonstrated earlier. For comparison, the fitness score of each of these artifacts is, for artifacts produced from images A, B, and C respectively: 0.836, 0.762, and 0.860 with an average score

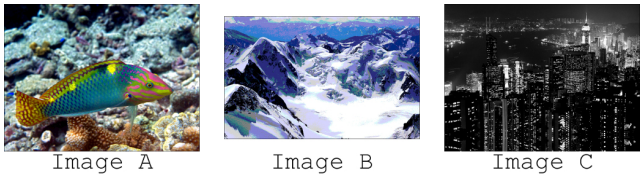

Figure 7: The most fit artifacts for each indicated source image discovered using the *similarity* fitness function.

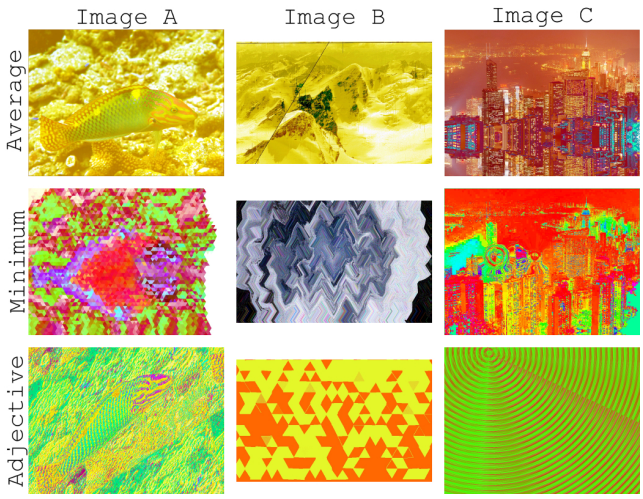

Figure 8: The most fit artifacts for each indicated source image and fitness function for the adjective 'happy'.

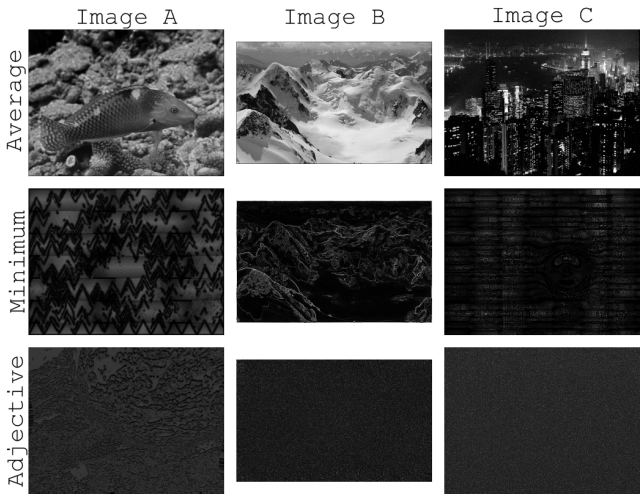

Figure 9: The most fit artifacts for each indicated source image and fitness function for the adjective 'sad'.

of 0.820. That being said, these artifacts are still quite close to the source images, and any resemblances to any of the specified adjectives are obviously happenstance.

For the *average* fitness function, arguably all three of the 'happy' images convey their adjective by applying bright colored filters (Figure 8). All three of the 'sad' images are made more sad by converting them to dark black and white images (Figure 9). Two out of the three 'fiery' images are fiery by primarily coloring with oranges and reds (Figure 10). Image B also looks bright and molten in texture, and some of the buildings in the background of image C almost look on fire. All three 'wet' images are debatably wet, mostly by implementing blue filters (Figure 11). Although, Image B actually looks like it is being viewed through a window soaked during a downpour. None of the 'peaceful' images look any more peaceful than their sources; and very little if anything has changed (Figure 12). With the odd exception of the 'peaceful' images, *average* does quite well at conveying adjectives; however, most of the images don't use much more than simple color filters to do so. In our estimation, for the *average* artifacts, 'happy' B and C, 'fiery' B and C, and 'wet' B satisfy the objectives for object rendering as outlined earlier.

For the *minimum* fitness function, two of the 'happy' images, A and C, are made happy by incorporating many bright colors. Image A looks kaleidoscopic and image C has some rainbow effects. Image B seems out of place, though close inspection will reveal that it may have received a high fitness because of many bright colors as well. While perhaps difficult to notice at first, both image A and B maintain the presence of the source image. All of the 'sad' images are quite dark, suggesting sadness. Image A and C may look like they have eliminated the source images, but the vague shape of the fish is visible within the squiggles of image A, and close inspection of image C will reveal many of the city lights behind the heavy distortion. The three 'fiery' images could be considered 'fiery'. Image A literally looks on fire and image C looks molten. All three 'wet' images appear wet; as with *average*, this is primarily accomplished by making the images blue. Image B does look like the image is now reflected off of a lake, and image C is a bit bleary and wavy giving it ever so slightly the look of being underwater. With the exception of image A, the 'peaceful' images aren't even recognizable, nor do they look peaceful in the way 'peaceful' is reflected in the training images. We're beginning to get a sense of how DARCI interprets 'peaceful' though. In our estimation, of the *minimum* images, 'happy' A and C, all 'sad' and 'fiery' images, and 'wet' B and C satisfy the objectives for object rendering. While 'happy' B and 'peaceful' A are *interesting* representations of the source image, they do not convey the adjective properly.

In the case of the *adjective* fitness function, we see that with three exceptions ('happy' A, 'sad' A, and peaceful 'C'), the source image is undetectable. 'Happy' A and 'sad' A do fit their adjectives, but 'peaceful' C does not. Interestingly, in our estimation *adjective* does not depict the given adjectives as well as *average* or *minimum*. This can be attributed in part to the system exploiting the VLA's neural networks with extreme and unnatural image features.

With all three of these fitness functions, we have seen unsatisfactory performance with 'peaceful'. However, this poor performance goes beyond DARCI's strange interpretation of what makes an image 'peaceful' (apparently being purple and noisy). That can be attributed to inadequate learning by the VLA, perhaps because of limited available training data. One could even make the case for it being a creative expression of 'peaceful'. The other problem here is the fact that for 'peaceful' artifacts, the three *average* artifacts were virtually unmodified from the source image, and that two of the *minimum* artifacts completely obfuscated the source image. This issue can be explained by a problematic interaction between the similarity and adjective matching metrics for 'peaceful'.

The 'peaceful' neural network output has very low vari-

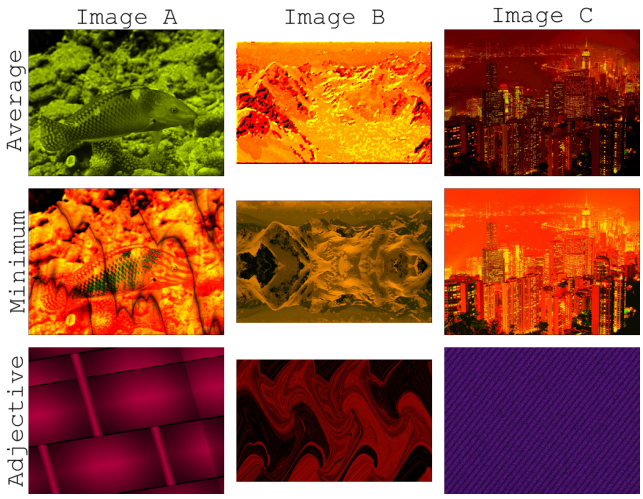

Figure 10: The most fit artifacts for each indicated source image and fitness function for the adjective 'fiery'.

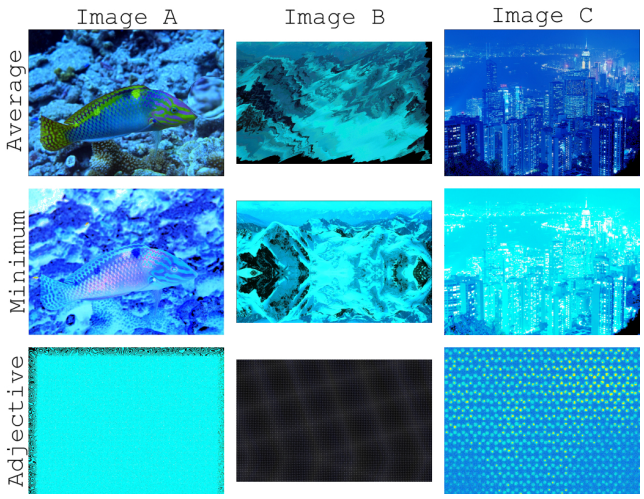

Figure 11: The most fit artifacts for each indicated source image and fitness function for the adjective 'wet'.

ance compared to the other neural networks, and a mean slightly under 0.5. The variance is so low that the highest 'peaceful' neural network outputs encountered are not much higher than the lowest similarity score possible (0.5). Thus, the *minimum* fitness function is effectively acting like the *adjective* fitness function for 'peaceful'. In the case of *average*, the variance is so low that the smallest changes in similarity still overshadow any changes in adjective matching. This example illustrates that despite our best efforts to balance the two metrics, incongruities between the two can still occur. Thus, for future work, a dynamic solution that takes into consideration certain statistics about each metric may be in order.

Selection After Last Shift

As indicated earlier, the *alternate* and *converge* fitness functions need a different selection method than that used above. As suspected, using most-fit selection resulted in artifacts that were either similar to those in Figure 7 or completely abstract like the images produced with *adjective*. The assumption with *alternate* and *converge* is that even though only a single metric is in effect at each generation, the genetic algorithm will not be able to converge to either because of constant shifts in the metric, and will instead find an interesting and unexpected solution.

With this in mind, the selection criteria that we use here is to pick the most fit artifact from the last *shift* in metric. This is the point at which we would expect to find the most surprising artifacts. We define a *shift* in the metric as the changing from the similarity metric to the adjective matching metric or vice versa. For *alternate* this is the shift from similarity to adjective matching at generation 100 which we will call *alternate-adjective*, and for *converge* it is the shift from adjective matching to similarity also at generation 100 which we will call *converge-similarity*. Since the direction of the shift may strongly effect the outcome, we have also selected the most fit artifact from generation 99 for *alternate* (adjective matching to similarity) and generation 80 for *converge* (similarity to adjective matching). We will call these two approaches respectively *alternate-similarity* and *converge-adjective*.

The results of these experiments are in Figures 13 to 15. In the interest of space, we do curate these images by only showing those artifacts that are neither over nor under-filtered (i.e. *interesting*) based on observations similar to those made for the earlier experiments. In the case of *alternate-similarity*, there were no artifacts produced that weren't under-filtered. Most had tinting or small distortions, but none were *interesting*.

Figure 13 shows *interesting* artifacts that were selected with *alternate-adjective*. This particular fitness function and selection criteria yielded the most numerous *interesting* artifacts of the four configurations. In this case, all but one of the not-shown artifacts were too abstract. Of the remaining *interesting* artifacts, all but the unusual 'peaceful' images arguably convey the intended adjectives.

Next, Figure 14 shows *interesting* artifacts selected with *converge-adjective*. Most of the other artifacts selected obfuscated the source image too much. Here, with the exception of 'fiery' A and perhaps 'fiery' B, the images convey the intended adjectives.

Finally, Figure 15 shows the *interesting* artifacts selected with *converge-similarity*. While the images shown are adequately *interesting*, we don't consider them as distinguished as those in the previous two examples. All of the other artifacts were too similar to the source image to warrant display. All of the displayed artifacts do convey the given adjectives.

Filter Sequence Length

Functionally, much of the quality of an artifact can be attributed to the length of the artifact's genotype. The genotype is the "genetic" encoding of the artifact, and in the image rendering subsystem is a sequence of image filters. The more filters used to render a source image, the more likely

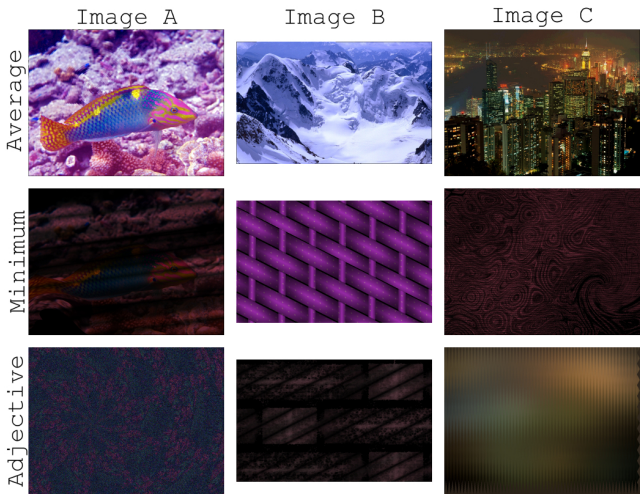

Figure 12: The most fit artifacts for each indicated source image and fitness function for the adjective 'peaceful'.

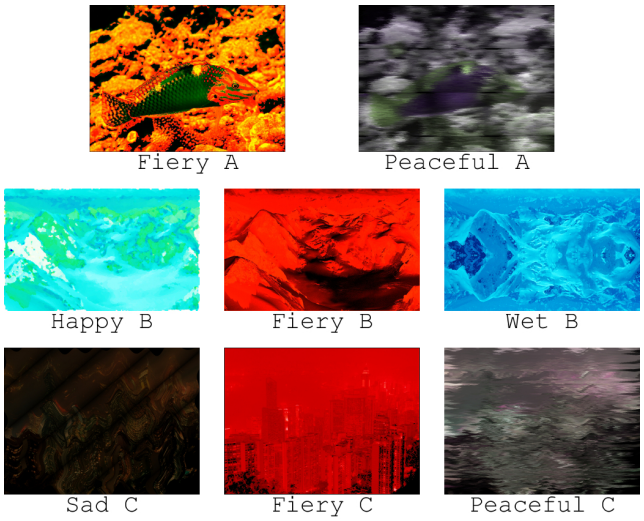

Figure 13: Artifacts selected for the indicated source images and adjectives for the *alternate-adjective* fitness function.

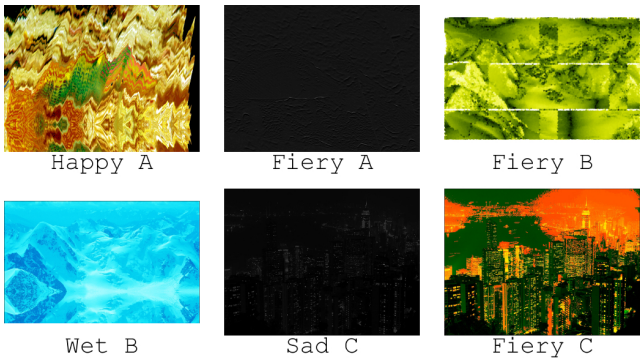

Figure 14: Artifacts selected for the indicated source images and adjectives for the *converge-adjective* fitness function.

the artifact will become abstract. The fewer filters used, the more likely the artifact will not deviate from the source image. Figure 16 shows the average genotype length (in number of filters) for each fitness function explored in this paper over the 100 epochs of evolution. The top performing fitness functions show a comfortable balance between too many and too few filters. *Minimum* does this the best.

Conclusions

The motivation behind this work has been to improve DARCI's ability to independently curate its own artifacts. All of the artifacts displayed in this paper were fully curated by DARCI under various selection criteria, with only a few indicated exceptions for space.

We show that DARCI is autonomously able to consistently create and select images that reflect the requested adjective with four out of five adjectives. This demonstrates the quality of the new adjective matching metric. We also demonstrate that the similarity metric functions as intended.

We explored a variety of fitness functions combining two metrics with varying degrees of success. Each method of combining the metrics had its own biases but, from our analysis, the *minimum* fitness function performed the best. Over half of the artifacts selected with this fitness function satisfied the goals of the image rendering subsystem—arguably a significant step in decreasing the latent heat effect in DARCI. We attribute the success of *minimum* to the fact that it allows the genetic algorithm to naturally shift evolutionary focus to the metric that is suffering the most.

We are confident that the improvements made to the image rendering subsystem in this paper will significantly decrease the latent heat effect in DARCI. We intend to test this theory in the future by conducting a thorough online survey comparing this improved version of DARCI to other versions, and perhaps even to humans. To further improve the image rendering subsystem described in this paper, we also intend to pursue more adaptable variations of the metrics outlined here. Metrics that will adapt their output in response to the features of other metrics.

References

Baluja, S.; Pomerleau, D.; and Jochem, T. 1994. Towards automated artificial evolution for computer-generated images. *Connection Science* 6:325–354.

Bay, H.; Ess, A.; Tuytelaars, T.; and Gool, L. V. 2008. Speeded-up robust features (SURF). *Computer Vision and Image Understanding* 110:346–359.

Boden, M. A. 1999. *Handbook of Creativity*. Press Syndicate of the University of Cambridge. chapter 18.

Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms (second edition)*. Routledge.

Colton, S., and Wiggins, G. A. 2012. Computational creativity: The final frontier. In *20th European Conference on Artificial Intelligence*, 21–26.

Colton, S. 2008. Creativity versus the perception of creativ-

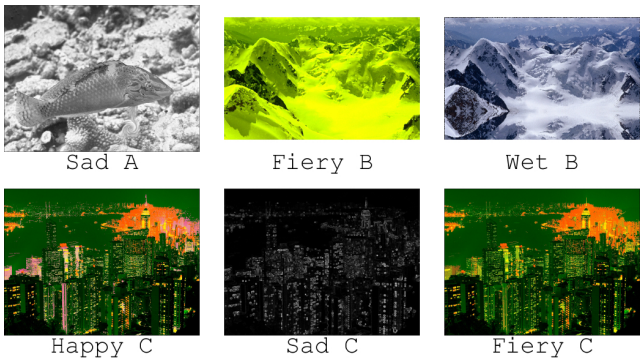

Figure 15: Artifacts selected for the indicated source images and adjectives for the *converge-similarity* fitness function.

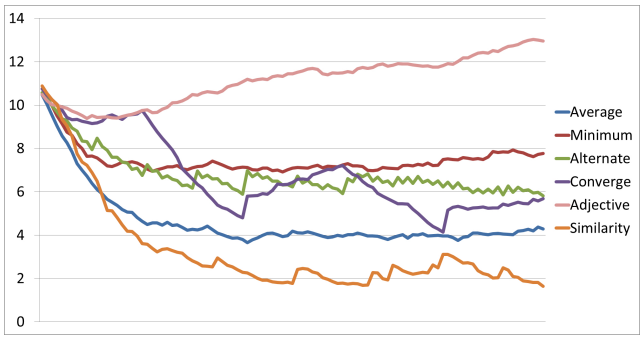

Figure 16: The average lengths of genotypes across all fitness functions over 100 epochs of evolution.

ity in computational systems. *Creative Intelligent Systems: Papers from the AAAI Spring Symposium* 14–20.

Csurka, G.; Dance, C. R.; Fan, L.; Willamowski, J.; and Bray, C. 2004. Visual categorization with bags of keypoints. In *Proceedings of the Workshop on Statistical Learning in Computer Vision*, 1–22.

Datta, R.; Joshi, D.; Li, J.; and Wang, J. Z. 2006. Studying aesthetics in photographic images using a computational approach. *Lecture Notes in Computer Science* 3953:288–301.

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

Elkan, C. 2003. Using the triangle inequality to accelerate *k*-means. In *Proceedings of the Twentieth International Conference on Machine Learning*, 147–153.

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., and Smeulders, A. 2000. Combining color and shape invariant features for image retrieval. *IEEE Transactions on Image Processing* 9:102–119.

Heath, D., and Norton, D. 2009. DARCI (Digital ARtist Communicating Intention). http://darci.cs.byu.edu.

Heath, D.; Norton, D.; and Ventura, D. 2013. Autonomously communicating conceptual knowledge through visual art. In *Proceedings of the 4th International Conference on Computational Creativity*, 97–104.

King, I.; Ng, C. H.; and 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., and Chen, T. 2009. Aesthetic visual quality assessment of paintings. *IEEE Journal of Selected Topics in Signal Processing* 3:236–252.

Machado, P.; Romero, J.; and Manaris, B. 2007. Experiments in computational aesthetics: An iterative approach to stylistic change in evolutionary art. In Romero, J., and Machado, P., eds., *The Art of Artificial Evolution: A Handbook on Evolutionary Art and Music*. Berlin: Springer. 381–415.

Maher, M. L.; Brady, K.; and Fisher, D. H. 2013. Computational models of surprise as a mechanism for evaluating creative design. In *Proceedings of the 4th International Conference on Computational Creativity*, 147–151.

Maher, M. L. 2010. Evaluating creativity in humans, computers, and collectively intelligent systems. In *DESIRE '10 Proceedings of the 1st DESIRE Network Conference on Creativity and Innovation in Design*, 22–28.

Norton, D.; Heath, D.; and Ventura, D. 2010. Establishing appreciation in a creative system. In *Proceedings of the 1st International Conference on Computational Creativity*, 26–35.

Norton, D.; Heath, D.; and Ventura, D. 2011a. An artistic dialogue with the artificial. In *Proceedings of the 8th ACM Conference on Creativity and Cognition*, 31–40. New York, NY, USA: ACM.

Norton, D.; Heath, D.; and Ventura, D. 2011b. Autonomously creating quality images. In *Proceedings of the 2nd International Conference on Computational Creativity*, 10–15.

Norton, D.; Heath, D.; and Ventura, D. 2013. Finding creativity in an artificial artist. *Journal of Creative Behavior* 47(2):106–124.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17:67–99.

Sivic, J.; Russell, B. C.; Efros, A. A.; Zisserman, A.; and Freeman, W. T. 2005. Discovering objects and their location in images. *International Journal of Computer Vision* 1:370–377.

Wang, W.-N.; Yu, Y.-L.; and 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.

Zujovic, J.; Gandy, L.; and Friedman, S. 2007. Identifying painting genre using neural networks. *miscellaneous*.