Conveying Semantics through Visual Metaphor

DERRALL HEATH, Brigham Young University
DAVID NORTON, Brigham Young University
DAN VENTURA, Brigham Young University

In the field of visual art, metaphor is a way to communicate meaning to the viewer. We present a computational system for communicating visual metaphor that can identify adjectives for describing an image based on a low-level visual feature representation of the image. We show that the system can use this visual-linguistic association to render source images that convey the meaning of adjectives in a way consistent with human understanding. Our conclusions are based on a detailed analysis of how the system's artifacts cluster, how these clusters correspond to the semantic relationships of adjectives as documented in WordNet, and how these clusters correspond to human opinion.

Categories and Subject Descriptors: []

General Terms: Design, Algorithms, Performance

Additional Key Words and Phrases: visual metaphor, evolutionary art, neural networks, clustering

1. INTRODUCTION

Metaphor can be a powerful method for communication. Metaphors are usually associated with linguistics and research has been done in the AI community exploring the interpretation and generation of metaphor within the linguistics domain [Veale and Hao 2007; Veale and Li 1991]. However, metaphor also exists in the visual domain. Traditionally, the term *visual metaphor* is used to mean the representation of a target concept by a different source concept [Forceville 1996]. The source replaces the target in the image and can thus convey a new interpretation of the source concept or become a symbol for it. For example, the concept of a dove is often used to represent the target of peace in images. In this example, peace is an abstract concept and a dove is concrete; but, this does not always have to be the case. A visual metaphor can be as simple as using an icon of an envelope to indicate access to email on a computer terminal. Visual metaphor can be a powerful tool for conveying meaning in an image and has seen effective use in art and advertising [Brown 2011; Kaplan 2005].

We have developed a computer system, called DARCI (Digital ARtist Communicating Intention), that is designed to autonomously create images that convey meaning. The value of visual metaphor in such a system is apparent and is one of the design goals for this system. However, automatic incorporation of visual metaphor in an image is nontrivial, requiring a rather mature understanding of language as well as the ability to visually represent and recognize a word in an image. As a step in this direction, DARCI currently only works with adjectives. Adjectives, more easily than nouns and verbs, can be used to describe an image based on general features such as color distribution, lighting, and repeating patterns. DARCI is presently designed to render an image to communicate a list of specified adjectives using a variety of filters and other image processing techniques. In effect, DARCI can express the meaning of adjectives visually in the images it produces. Since DARCI cannot interpret or render nouns, implementing the traditional view of conveying visual metaphor by replacing a target with a source is severely limited. In a broader sense however, DARCI can

This work is supported by the National Science Foundation, under grant IIS-0856089.

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

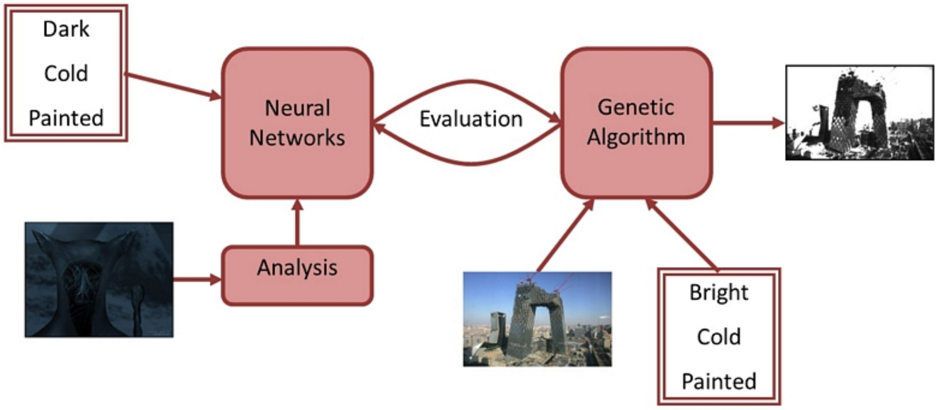


Fig. 1. Overview of DARCI's artifact creation process.

represent an adjective, the target, as a unique filter, an abstract source. This sense of visual metaphor does exist in human interaction. For example, the color red is often used to denote danger. For brevity, throughout this paper we will refer to this broader sense of visual metaphor as simply *visual metaphor*. Though limited, this is an important first step in the direction of conveying fully automated visual metaphor in the narrower, more traditional sense.

DARCI learns the meaning of adjectives, or rather adjective synonym sets—called synsets—by building associations between low-level visual features and the synsets themselves using a series of neural nets trained with human labeled data [Norton et al. 2010]. The process is augmented by taking advantage of the synset relationships found in WordNet [Fellbaum 1998]. DARCI renders images by applying various image filters to a source image (not to be confused with the source concept in visual metaphor). The filter settings are learned with an evolutionary mechanism that is governed by the visual-linguistic associations discovered by the neural nets [Norton et al. 2011]. Rather than the traditional approach to evolutionary art where humans evaluate the products of each generation by hand [Rooke 2002; Todd and Latham 1992; Sims 1991; Secretan et al. 2011], the fitness function is the output of these neural nets. Figure 1 outlines this process of creating artifacts. While there are other approaches using automated and dynamic fitness functions in evolutionary art [Oranchak 2007; Machado and Cardoso 2002; Greenfield 2002; DiPaola and Gabora 2009; Greenfield 2007; Baluja et al. 1994], none explore the communication of meaning as is done here with DARCI.

In previous work we have shown a degree of success in both labeling and generating images with respect to semantic content [Norton et al. 2010; 2011]. However, analyzing the meaning contained in a particular rendering is an inherently subjective task. Thus, we are interested in additional and more thorough approaches to analyzing the effectiveness of our system. Most of our previous evaluations of DARCI have required the opinion of human volunteers. Unfortunately, using human judges in an evaluation can be costly and highly variable. In this paper we explore an approach to evaluating the system that does not require human interaction. Instead this approach analyzes the way images produced by DARCI cluster and in turn how these cluster relationships correspond with synset relationships found in WordNet. In order to validate our approach, we compare an agglomerative clustering of DARCI's images by their features to how human volunteers' rankings of DARCI's images cluster. Each set of results provides evidence to support our claim that the system is capable of expressing visual metaphor in the artifacts it produces.

We begin this paper by providing a detailed overview of DARCI's algorithms as explained in previous work. We then introduce the methodology for our new evaluation metrics that are the focus of this paper. Next we describe the various experiments we ran with our evaluation methodology in mind. Here we also analyze both how effective our new evaluation metrics are and how effective

DARCI is at using visual metaphor to communicate the meaning of adjectives in images. Finally we draw conclusions and suggest future direction for this work.

2. SYSTEM OVERVIEW

In previous work we have described and evaluated the various algorithms that operate within DARCI [Norton et al. 2010; 2011]. For convenience, in this section we reproduce those details necessary to appreciate the results of this paper.

2.1. Visuo-Linguistic Association

In order for DARCI to make associations between images and their meaning, we present the system with images that are labeled appropriately. For now, we have reduced descriptive labels exclusively to delineated lists of adjectives. Also, since the raw feature space of a typical image is intractable, we have selected 102 real-valued low-level image features to extract from each image [Norton et al. 2010]. These features were selected based on prior research in the area of image feature extraction [Gevers and Smeulders 2000; Datta et al. 2006; Li and Chen 2009; Wang et al. 2006; King ; Wang and He 2008] and include general measures of color, light, texture, and shape [Norton et al. 2010]. Three of these features, for example, are the average RGB (red, green, and blue) values of each pixel in an image. Because the magnitude of these 102 features varies dramatically from feature to feature, we have standardized the image features using a core set of training images (approximately 2000 diverse images). Throughout this paper, when we refer to image features, we are referring to the standardized values.

We use WordNet's [Fellbaum 1998] database of adjective synsets to give us a large set of descriptive labels. Even though our potential labels are restricted to a single lexical category, the complete set of WordNet adjective synsets can allow for images to be described by their emotional effects, most of their aesthetic qualities, many of their possible associations, and even, to some extent, by their subject.

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, DARCI lists seven adjectives that it 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 image to synset associations is a *multi-label classification* problem [Tsoumakas and Katakis 2007], meaning each image can be associated with more than one synset. To handle multi-label classification, we use a collection of artificial neural networks (ANNs) that we call appreciation networks. There is an appreciation network 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 neural networks can be dynamically added to the collection to accommodate the new synsets. As of writing this paper, there are 211 appreciation networks. This means that DARCI essentially "knows" 211 synsets. The appreciation networks are trained using standard backpropagation and output a single real value, between 0 and 1, indicating the degree to which a given image can be described by the networks' corresponding synset.

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 use antonym relationships found in WordNet in addition to statistical correlations discovered in our data as described in prior research [Norton et al. 2010].

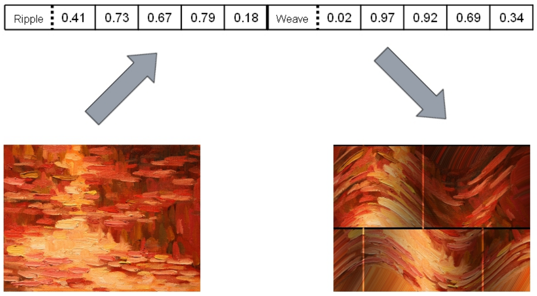


Fig. 2. Sample genotype (top) applied to a source image (left) resulting in the phenotype (right). The genotype is a list of image filters with parameters. “Ripple” and “Weave” are the names of two (of ninety-two) possible filters.

2.2. Image Generation

DARCI uses an evolutionary mechanism to render images so that they visually express the meaning of given synsets. Because of the innate ability of evolution to yield novel solutions to problems, evolutionary methods are frequently used in generative art [Gero 1996].

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 a separate, persistent 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. The evolutionary mechanism for both modes functions as follows.

The genotypes that comprise each gene pool are lists of filters (and their accompanying parameters) for processing a source image. The processed image is the phenotype. 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 [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 2 gives an example of a genotype and its phenotype. 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 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).

Every generation of the evolutionary mechanism, each phenotype is created from 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 function used to evaluate the fitness of each phenotype created during the evolutionary process can be expressed by the following equation:

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

where f^P is the vector of 102 image features (see Section 2.1) for a given phenotype and $A : F^P \rightarrow [0, 1]$ and $I : F^P \rightarrow [0, 1]$ are two functions: appreciation and interest. These functions compute a real-valued score for a given phenotype (here, F^P represents the set of all phenotype feature vectors). $\lambda_A + \lambda_I = 1$, and for now, $\lambda_A = \lambda_I = 0.5$.

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

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

where C is the set of synsets to be portrayed, $\text{net}_w(\cdot)$ is the output of the appreciation network for synset w, and $\alpha_w = 1/|C|$ (though this can, of course, be changed to weight synsets unequally). As indicated, f^P, the feature vector of the phenotype, is the input to each appreciation network.

The interest function I penalizes phenotypes that are either too different from the source image, or are too similar. This can be expressed with the Equations 3 - 5 as follows:

$$n = \sum_i \sigma(f_i^S, f_i^P) \tag{3}$$

$$\sigma(f_i^S, f_i^P) = \begin{cases} 0, & |f_i^S - f_i^P| > 0.3 \\ 1, & |f_i^S - f_i^P| < 0.3 \end{cases} \tag{4}$$

where f_i^S represents feature i of the source image and f_i^P represents feature i of the phenotype. The similarity threshold of 0.3 was chosen empirically. Equations 3 and 4 give a count, n, of how many features between the phenotype and source image are similar. The interest score is calculated using n as follows:

$$I(f^P) = 1 - \begin{cases} \frac{\tau_d - n}{\tau_d}, & n < \tau_d \\ \frac{n - \tau_s}{|f| - \tau_s}, & n > \tau_s \\ 0, & \tau_d \leq n \leq \tau_s \end{cases} \tag{5}$$

where 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 (102). In effect, if there are between 20 and 57 image features that are similar between the source image and the phenotype in question (see Equation 3), then the interest score for the phenotype is 1—the maximum score. Otherwise the interest score will degrade according to Equation 5.

Our evolutionary mechanism for image generation operates similar to other standard genetic algorithms. As such, there are several more components to the mechanism that require further discussion including the phenotype selection process, crossover, mutation, and migration.

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

Table I. Parameters used for the evolutionary mechanism.

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

genotype are changed. The degree of this change, Δl_i, for each parameter, i, is determined by one of the following two equations chosen randomly with equal probability:

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

$$\Delta l_i = -l_i \cdot rand\left(0, \frac{(|l| + 1) - |\Delta l|}{|l|}\right) \tag{7}$$

Here, l_i is the value of parameter i prior to mutation, $|l|$ is the total number of parameters in the mutating filter, $|\Delta l|$ 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).

3. EVALUATION METHODOLOGY

In order to evaluate DARCI's artifacts more consistently and without direct human involvement, we use a clustering algorithm to find the inherent groupings of the images. DARCI's artifacts should cluster in ways comparable to relationship clusters found within WordNet.

WordNet is an extensive ontology of the English language consisting of over 117,000 synsets, the basic unit of WordNet, and their relationships. Since individual words can have different meanings, or senses, WordNet is not structured around the words themselves; rather, it is structured around these meanings. A synset is a collection of different words that share the same meaning and can be used interchangeably. For example "bright" and "smart" can mean the same thing and thus together form one synset. Another sense of "bright" means the same thing as a sense of "brilliant" and a sense of "vivid". These form another distinct synset. In WordNet, all synsets are placed into one of four part-of-speech categories: nouns, adjectives, verbs, and adverbs. This paper is primarily concerned with the more than 18,000 adjective synsets.

Many of the relationships documented in WordNet can provide us with a measure of semantic similarity between synsets. For adjectives, these relationships include *adjective clusters*, *satellites*,

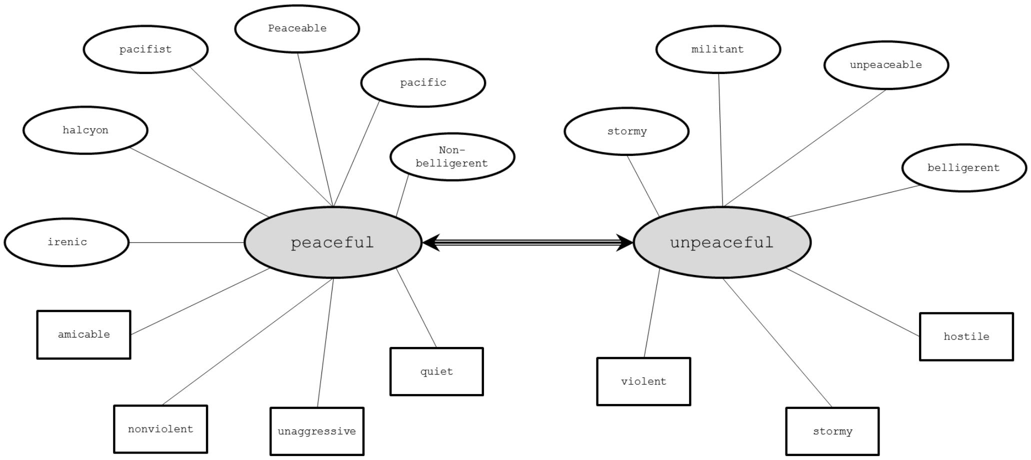


Fig. 3. The adjective cluster for the antonym pair "peaceful/unpeaceful" as contained in WordNet. "Peaceful" and "unpeaceful" are the head synsets. The synsets in ovals are satellites while the synsets in rectangles are related concepts (technically not part of the adjective cluster).

antonyms, and *related concepts*. Most adjective synsets are contained in adjective clusters which consist of two (rarely three) head synsets and their respective satellites. Antonyms are synsets that are opposite in meaning and are commonly associated. The head synsets in an adjective cluster are bound by an antonym relationship making the clusters polar groupings. The satellites of a cluster are synsets similar in meaning to their respective head. Satellites are indirectly antonymous to the opposing head satellites. Related concepts are synsets that do not belong to the same adjective cluster, but share some similarity in meaning. As an example, Figure 3 shows the adjective cluster for the antonym pair "peaceful" and "unpeaceful".

Ideally, DARCI's artifacts rendered with synsets that share the same satellite head should tend to cluster more frequently than artifacts rendered with synsets that are only related concepts, which should in turn cluster more frequently than artifacts rendered with more distantly related concepts or concepts that are altogether unrelated, and so on. Unfortunately this ideal scenario is not wholly practical due both to limitations inherent in WordNet, and to the complex nature of meaning in images.

Since WordNet's relationships are determined through a formal analysis of the language by linguistic experts, they don't necessarily agree with colloquial interpretations. For example, in WordNet, "happy" and "sad" are *not* direct antonyms. The antonym of "sad" is "glad", and the antonym of "happy" is "unhappy" (in a distinct adjective cluster from "sad"). "Sad" and "happy" do share related concepts that are antonyms, so there is an indirect connection between the two adjectives. Still, this demonstrates a slight incongruity between commonly assumed meanings (comprising DARCI's hand-labeled training dataset) and meanings defined in WordNet. Furthermore, WordNet's related concepts are arguably not complete as illustrated by the fact that they are not all bidirectional. For example, "glad" is a related concept of "happy", but "happy" is not a related concept of "glad". Finally, certain terms that are technically semantically unrelated, do have semantic connections in people's minds. For example "warm" as in the temperature has a distinctly different meaning than "warm" as in the use of color; however, people relate the two because one suggests the other. WordNet does not usually include these connections.

Despite these limitations, there is value in using WordNet to evaluate the way in which DARCI's artifacts are clustered. Having clusters that disagree with WordNet does not necessarily indicate a failure in DARCI since arguably WordNet is not complete; however, agreeing with WordNet *does* signify a degree of success since WordNet's relationships are widely accepted, giving us a somewhat objective measure of a quality that is inherently subjective.

We use the EM (Expectation Maximization) algorithm found in Weka [Hall et al. 2009] to cluster the images with the same 102 low-level image features used to inform DARCI's neural networks. These features include general measures of color, light, texture, and shape [Norton et al. 2010]. For these experiments, we specify the number of clusters to be equal to the number of adjectives that are present among the images.

We perform two sets of experiments. In the first set, we analyze how the difference in clustering quality between different groups of artifacts illustrates the expression of meaning in the artifacts. In the second set of experiments, we show how the sequence in which images agglomeratively cluster demonstrates the expression of meaning. In both sets of experiments, DARCI produces images to be clustered using the evolutionary mechanism detailed earlier and in prior work [Norton et al. 2011].

4. RESULTS

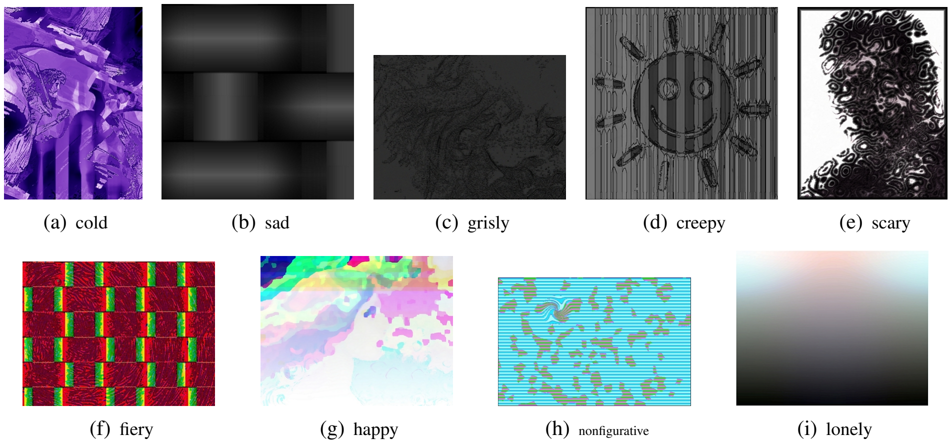
We perform two groups of experiments to explore clustering as an evaluation of the expression of semantics. In the first group, we analyze how DARCI's artifacts cluster into specified sets of adjective synsets. We explore different sets of synsets and examine the quality of the resulting clusters. In the second group of experiments, we use agglomerative clustering to build a hierarchy of DARCI's visual metaphors using two different sets of features as data points. Finally, in order to validate the agglomerative clustering results, we survey human volunteers by having them rank image similarity.

4.1. Cluster Quality

For our first series of experiments we selected two sets of synsets, a *distinct* set and a *similar* set. The distinct set contains five synsets that are either antonymous or conceptually unrelated to each other according to WordNet: "cold", "fiery", "happy", "nonfigurative" (as in abstract or nonfigurative art), and "sad." The similar set contains five synsets that are either conceptually related according to WordNet, or similar in mood to one another: "creepy", "grisly", "lonely", "sad", and "scary." Both sets contain the same synset for "sad." We began by creating a practice gene pool for each synset belonging to these sets. We trained these gene pools for 100 generations and then collected the most fit images created from the last 20 epochs of training. Example images from each of the nine synsets are shown in Figure 4; the source images DARCI used to generate these images are shown in Figure 5. We then performed EM clustering on each set's collection of images. Finally, we analyzed the resulting clusters.

Often clusters favored specific synsets; however, as there are variable source images, and many ways for synsets to be legitimately expressed with DARCI's rendering tools, clusters were usually composites of multiple synsets. It is difficult to draw conclusions from these experiments by simply looking at the breakdown of each cluster. However, we can see how the meaning of synsets is present in DARCI's artifacts by analyzing metrics obtained for each *synset*. We determined the F1 measure, precision, and recall for each synset in each cluster, and then isolated the best value of each metric for the various synsets. The cluster that the *best value* comes from tells us which cluster best represents each synset. The value itself tells us how well the synset is represented by that cluster. Additionally, we calculated the average cluster entropy and average cluster purity to determine how well the clustering matches the original adjective groupings. We repeated this clustering experiment five times using newly created gene pools each time. Tables II and III show the best F1 measure, precision, and recall, as well as average entropy and average purity, for the distinct and similar sets of synsets averaged across the five experiments.

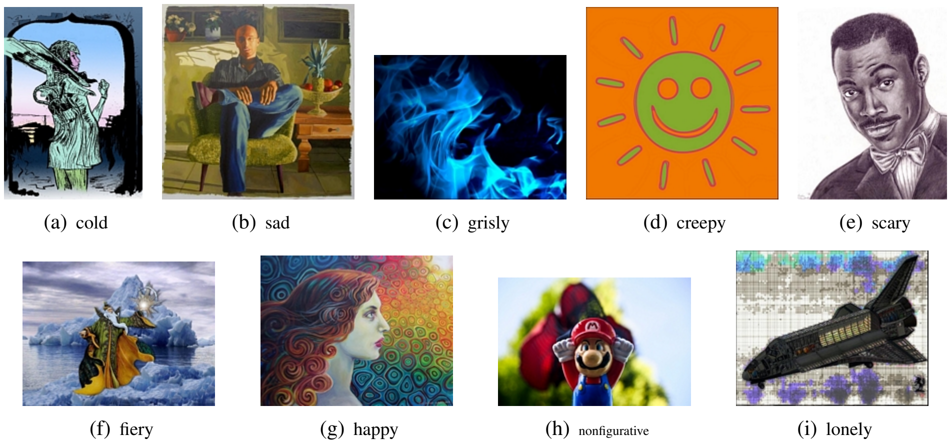
To help understand the results, consider as an example the adjective "cold" in Table II. The recall metric tells us that 57% of the "cold" images clustered together into one cluster (this is the largest grouping of "cold"). The precision metric tells us that in that cluster, 64.5% of the images were "cold." The F1 measure is essentially the weighted average of precision and recall. These metrics give us an overview of how well the images for each individual synset clustered. The metrics entropy



(a) cold (b) sad (c) grisly (d) creepy (e) scary

(f) fiery (g) happy (h) nonfigurative (i) lonely

Fig. 4. Examples of images created by DARCI during practice. These are created from the corresponding source images found in Figure 5



(a) cold (b) sad (c) grisly (d) creepy (e) scary

(f) fiery (g) happy (h) nonfigurative (i) lonely

Fig. 5. The source images DARCI used to generate the corresponding images in Figure 4.

and purity tell us how well, as a whole, all the images clustered. For all the metrics (except for entropy), the higher the value the better. It is clear that the distinct synsets have consistently better scores than the similar synsets. This means that more confusion is occurring in the synsets that are semantically more similar—the behavior we would expect in a system that is expressing appropriate meaning in its artifacts.

We performed the same experiment using commissioned images rather than practice images. Using the first of the practice gene pools for each synset just created, we commissioned DARCI with three source images for each synset. Each commission was given 100 epochs to develop. The top 40 images from each commission were collected making a total of 120 images per synset. The empirical results of clustering these commissions are shown in Tables IV and V. Again, the distinct synsets were more effectively clustered than the similar synsets as we would expect. It is interesting to note that "nonfigurative" in Table IV has a value of 1.000 for precision. This tells us that there was a cluster where 100% of images it contained were "nonfigurative." However, the recall metric

Table II. Best F1 measure, precision, and recall for each synset, as well as average entropy and average purity, after performing EM clustering on practice images for distinct synsets. (Lower is better for entropy.)

	F1	Precision	Recall
cold	0.565	0.645	0.570
fiery	0.569	0.614	0.590
happy	0.651	0.578	0.760
nonfigurative	0.520	0.555	0.540
sad	0.660	0.705	0.660
AVERAGES	0.593	0.619	0.624
ENTROPY	0.567	PURITY	0.584

Table III. Best F1 measure, precision, and recall for each synset, as well as average entropy and average purity, after performing EM clustering on practice images for similar synsets. (Lower is better for entropy.)

	F1	Precision	Recall
creepy	0.398	0.458	0.490
grisly	0.409	0.377	0.480
lonely	0.448	0.708	0.460
sad	0.481	0.463	0.540
scary	0.557	0.538	0.600
AVERAGES	0.459	0.509	0.514
ENTROPY	0.752	PURITY	0.458

Table IV. Best F1 measure, precision, and recall for each synset, as well as average entropy and average purity, after performing EM clustering on commissioned images for distinct synsets. (Lower is better for entropy.)

	F1	Precision	Recall
cold	0.784	0.987	0.650
fiery	0.429	0.362	0.525
happy	0.690	0.736	0.650
nonfigurative	0.717	1.000	0.558
sad	0.660	0.557	0.808
AVERAGES	0.656	0.729	0.638
ENTROPY	0.521	PURITY	0.623

Table V. Best F1 measure, precision, and recall for each synset, as well as average entropy and average purity, after performing EM clustering on commissioned images for similar synsets. (Lower is better for entropy.)

	F1	Precision	Recall
creepy	0.455	0.623	0.450
grisly	0.272	0.500	0.508
lonely	0.469	0.319	0.883
sad	0.429	0.292	0.808
scary	0.711	0.932	0.575
AVERAGES	0.467	0.533	0.645
ENTROPY	0.779	PURITY	0.445

then tells us that the cluster only contained 55.8% of the "nonfigurative" images. This indicates that just over half of the "nonfigurative" images were particularly distinct.

These results verify those obtained from clustering the practice images. Comparing the results of the commissioned images and the practice images, we see that the commissioned images have better values than practice. This is likely because we started each commission with the practice gene pools for each synset, which essentially gave the commissioned images a head start.

In preparation for the agglomerative clustering experiments, we added eleven synsets to the nine outlined above (see Figure 7), and commissioned DARCI with the same three images for each synset (again 100 epochs). An example commission from each of these 20 synsets for one of the source images is shown in Figure 6. It should be noted that the adjective "abstract" in this case means "abstract concept" and is thus in a slightly different synset than "nonfigurative." The synsets were selected to cover a spectrum of meanings with varying degrees of similarity while also being well represented in DARCI's training database. We then determined the number of WordNet related concepts or antonymous relationships between each pair of synsets. We call this value the *semantic distance* and represent positive relationships with a positive value and negative relationships (antonymous relationships) with a negative value. This distance is defined to be a value of 1 for synset pairs that share the same satellite head, and increases in magnitude by 1 for every relationship link (related concept or antonym) crossed. In addition, every time an antonym link is crossed, the distance changes sign. Table VI shows all of the semantic distances between pairs of synsets with a distance magnitude less than 10.

We took each pair of synsets listed in Table VI, and performed EM clustering over the two synsets' images. We then calculated the F1 measure, precision, and recall for each synset in each pair, as well as the average entropy and average purity for each pair. Finally, we averaged the metrics across the similar (positive semantic distance) and distinct (negative semantic distance) pairs of synsets. The results are recorded in Table VII. Yet again, the distinct pairs of synsets clustered more effectively than the similar pairs of synsets.

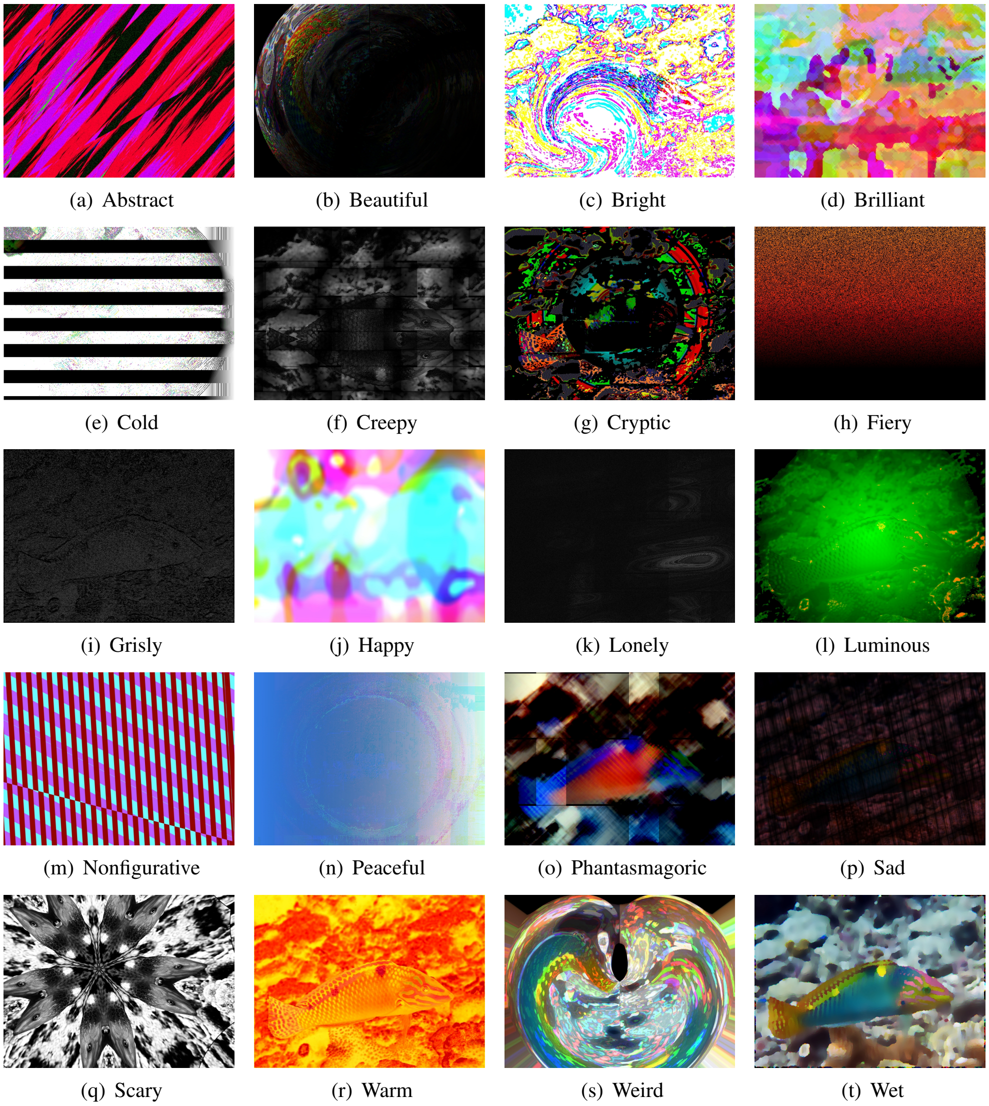

(a) Abstract (b) Beautiful (c) Bright (d) Brilliant
(e) Cold (f) Creepy (g) Cryptic (h) Fiery
(i) Grisly (j) Happy (k) Lonely (l) Luminous
(m) Nonfigurative (n) Peaceful (o) Phantasmagoric (p) Sad
(q) Scary (r) Warm (s) Weird (t) Wet

Fig. 6. A sample of commissioned results for one source image used in the clustering experiments.

4.2. Agglomerative Clustering

In these experiments we use agglomerative clustering based on the EM clustering algorithm and centroids to see if related adjectives will group together before grouping with unrelated adjectives. We performed agglomerative clustering on the 20 adjectives listed in Figure 7 in two ways: using all 102 image features, and simplifying the feature space to just our 12 color and lighting features. The original feature set is comprehensive, dealing with various aspects of color, lighting, texture, and shape within an image. For comparison, the smaller set (of 12) uses only color and lighting features, as they are predominately considered the most significant in image comparison research [Choras 2007]. To perform agglomerative clustering, we first cluster the images into the 20 adjective groups in same way as done in Section 4.1. We then calculate the centroid of each resulting cluster. Note that the centroid of each cluster is the average feature vector of all the images in each cluster.

Table VI. A list of all synset pairs with a semantic distance magnitude less than 10. Synsets are ranked from most similar to most opposite. Note that lower magnitude negative distance indicates a stronger antonymous relationship.

bright	luminous	1	bright	peaceful	7
grisly	scary	1	creepy	phantasmagoric	9
abstract	nonfigurative	2	bright	creepy	-9
abstract	phantasmagoric	5	creepy	luminous	-9
beautiful	bright	6	beautiful	creepy	-5
beautiful	luminous	6	creepy	peaceful	-5
nonfigurative	phantasmagoric	6	happy	sad	-3
peaceful	luminous	7	cold	fiery	-2

Table VII. The average F1 measure, precision, recall, entropy, and purity for similar pairs of synsets and distinct pairs of synsets when binary clustering is applied. (Lower is better for entropy.)

	F1	Precision	Recall	Entropy	Purity
similar	0.726	0.778	0.790	0.763	0.727
distinct	0.803	0.823	0.859	0.596	0.793

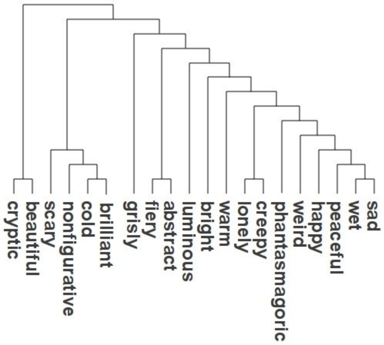


Fig. 7. Results of agglomerative clustering with 20 adjectives using all 102 images features.

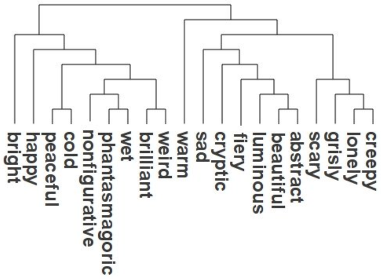


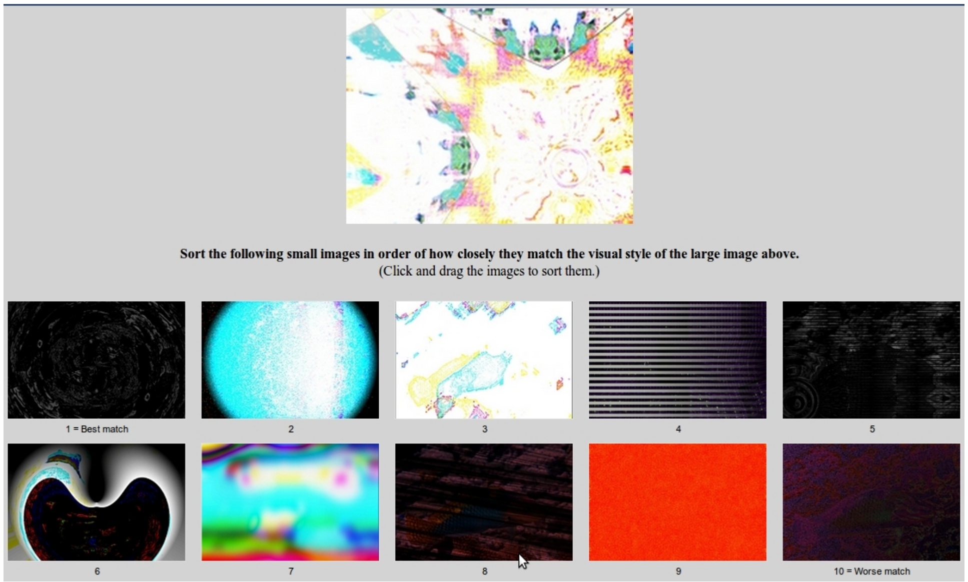
Fig. 8. Results of agglomerative clustering with 20 adjectives using only 12 color features.

We then decrement the number of clusters by one and recluster the new centroids. We repeat this process, decrementing the number of clusters each time, until there are only two clusters.

The results for each clustering can be seen in Figure 7 and Figure 8. With few exceptions (like "creepy" and "lonely"), the results of using all 102 features do not agree with the relationships in WordNet, nor with intuition. The results from using only the color and lighting features are considerably better. It is interesting to note that of the five semantically similar adjectives used in the previous experiment, four of them ("creepy", "lonely", "grisly", and "scary") clustered together before clustering with any other adjective. While the fifth one ("sad"), clustered with them before a majority of others. In contrast, the adjectives from the previous experiment that were semantically dissimilar ("sad", "happy", "nonfigurative", "fiery", and "cold"), remain distinct far up the cluster hierarchy.

4.3. Human Survey

To validate the clustering experiments just described, as well as to further evaluate DARCI's artifacts, we conducted a human survey. In this survey, we used the same images produced by DARCI that were used in the agglomerative clustering experiment for one of the source images. However, we limited the survey to only 10 of the 20 adjectives to avoid overburdening the users. The adjectives



Sort the following small images in order of how closely they match the visual style of the large image above.
(Click and drag the images to sort them.)

Fig. 9. An example screenshot of the human survey. The larger image on the top corresponds to a particular adjective. Using a drag and drop interface, the ten smaller images below must be sorted according to how well they match the visual style of the image above.

in the survey were “happy”, “sad”, “cold”, “bright”, “scary”, “creepy”, “weird”, “peaceful”, “luminous”, and “warm.” Recall that there are 40 images produced for each adjective bringing the total to 400 images. In the survey we presented each volunteer with one of DARCI’s 400 images randomly chosen. This image acted as a representative for one of the 10 adjectives of interest. Underneath this target image volunteers were shown additional random images (from these 400) for each of the 10 adjectives (in random order). There was no indication of which adjective went with which image. Using a drag and drop interface, the volunteers were required to sort the 10 images below based on how well they thought these images matched the target image above. The user repeated this process multiple times, each time with randomly chosen images. An example screen-shot can be seen in Figure 9.

There were 70 people that participated in the survey, each completing rankings for, on average, 8.3 target images for a total of 582 entries. Thus, each of the 10 adjectives had an average of 58 entries where each entry consisted of a target image (corresponding with the adjective) and 10 ranked images ranked according to their similarity with the target image. For each adjective we averaged the rankings of the 10 subordinate adjectives (1 being the closet match and 10 being the worst) to determine, overall, how similar each adjective’s images were to each other. The results can be seen in Table VIII.

The first thing to notice is that all 10 adjectives are most closely matched with themselves. This tells us that DARCI is consistent in rendering images that convey each adjective. It also tells us that DARCI’s representation of each adjective is distinct from other adjectives. We can also see that adjectives with similar meaning, such as “scary” and “creepy”, were consistently ranked closely to each other. Conversely, adjectives with dissimilar meaning, such as “happy” and “sad”, were consistently ranked farther from each other.

To compare the results of the survey with our clustering results, we performed agglomerative clustering on the survey data by using the average ranking as the distance metric between each adjective. We also reran the agglomerative clustering algorithm on these same 10 adjectives using

Table VIII. Results of the human survey. The adjective lists in the right column indicate the overall ranking of similarity between images rendered with the indicated adjectives and images rendered with the adjective in the left column.

Adjective	Ranking (left to right)
bright	bright, happy, peaceful, warm, scary, weird, sad, luminous, cold, creepy
cold	cold, sad, creepy, scary, peaceful, weird, bright, warm, happy, luminous
creepy	creepy, scary, sad, luminous, cold, warm, peaceful, weird, bright, happy
happy	happy, peaceful, weird, bright, scary, luminous, warm, sad, cold, creepy
luminous	luminous, warm, scary, creepy, happy, peaceful, bright, sad, weird, cold
peaceful	peaceful, weird, happy, scary, sad, creepy, luminous, bright, cold, warm
sad	sad, cold, creepy, scary, peaceful, weird, warm, happy, bright, luminous
scary	scary, creepy, luminous, weird, peaceful, sad, bright, happy, cold, warm
warm	warm, luminous, bright, sad, happy, scary, peaceful, cold, weird, creepy
weird	weird, scary, peaceful, sad, happy, cold, creepy, luminous, bright, warm

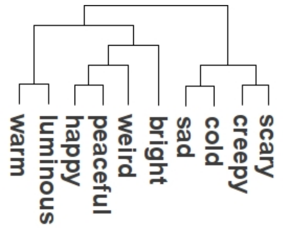


Fig. 10. Results of agglomerative clustering with 10 adjectives using the human survey data.

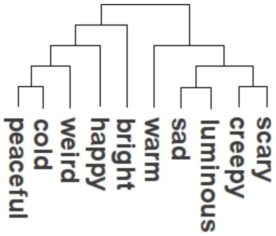


Fig. 11. Results of agglomerative clustering with 10 adjectives using the 12 color and lighting image features.

the 12 color and lighting features. The results for each clustering can be seen in Figure 10 and Figure 11.

With the exception of “cold”, “warm”, and “luminous”, the two clusterings are very similar in how they cluster. In both cases, we can see that similar adjectives, like “scary”, “creepy”, and “sad”, grouped together first. Likewise, we can also see that dissimilar adjectives, like “happy/sad” and “warm/cold”, grouped far apart. This not only validates DARCI’s ability to convey distinct meaning in its artwork, but also validates the use of clustering algorithms as an objective metric that can be used to evaluate DARCI’s artifacts. Note that in prior work we have done other human surveys that evaluate how well images created by DARCI actually communicate the meaning of a particular adjective compared to images created by human artists [Norton et al. 2011].

5. CONCLUSIONS

We have evaluated a system, DARCI, for using visual metaphor to communicate the meaning of specified adjectives in the images it renders. The system learns how to label images with adjectives and, in turn, to render them appropriately by training on eclectic human-labeled data. We have shown that DARCI can produce artifacts that will cluster in ways that reflect the supplied adjectives’ relationships according to WordNet. In addition, according to human opinion, DARCI can depict a distinct visual style for each synset the system renders. If we acknowledge that the meaning of words can be at least partially expressed by their relationships to each other, then our results indicate that DARCI is indeed learning how to visually convey the meaning of words. DARCI's ability to generate visual metaphor represents a unique stepping stone in the exploration of metaphor in AI.

We have also demonstrated a unique way to use clustering to evaluate the visual communication of meaning without direct human involvement. To validate this evaluation method, we have shown that it favorably compares with human opinion. Such a metric may be useful in any research interested in the empirical analysis of visual images.

In the future, we will explore the possibility of generating visual metaphor with nouns in addition to adjectives allowing for a more complete use of visual metaphor. This will require the development of higher-level feature extraction methods when building visual-linguistic associations. Further, adding nouns to DARCIs vocabulary suggests a close relationship to content-based image retrieval research and may allow us to leverage techniques from that field.

REFERENCES

BALUJA, S., POMERLEAU, D., AND JOCHEM, T. 1994. Towards automated artificial evolution for computer-generated images. *Connection Science 6*, 325–354.

BROWN, B. 2011. *Cinematography: Theory and Practice*. Focal Press, Chapter 5, 67–75.

CHORAS, R. S. 2007. Image feature extraction techniques and their applications for cbir and biometrics systems. *International Journal of Biology and Biomedical Engineering 1*, 6–16.

COLTON, S., GOW, J., TORRES, P., AND CAIRNS, P. 2010. Experiments in objet trouvé browsing. *Proceedings of the 1st International Conference on Computational Creativity*, 238–247.

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.

FELLBAUM, C., Ed. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

FORCEVILLE, C. 1996. *Pictorial Metaphor in Advertising*. New York: Routledge.

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.

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

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

HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P., AND WITTEN, I. H. 2009. The weka data mining software: an update. *SIGKDD Explor. Newsl. 11*, 10–18.

KAPLAN, S. 2005. Visual metaphors in print advertising for fashion products. In *Handbook of Visual Communication: Theory, Methods, and Media*, K. Smith, S. Moriarty, G. Barbatsis, and K. Kenney, Eds. New Jersey: Lawrence Erlbaum Associates, Publishers, Chapter 11, 167–177.

KING, I. Distributed content-based visual information retrieval system on peer-to-pear(p2p) network. http://appsrv.cse.cuhk.edu.hk/~miplab/discovir/.

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. AND CARDOSO, A. 2002. All the truth about NEvAr. *Applied Intelligence, Special Issue on Creative Systems 16*, 101–119.

NORTON, D., HEATH, D., AND VENTURA, D. 2010. Establishing appreciation in a creative system. *Proceedings of the 1st International Conference on Computational Creativity*, 26–35.

NORTON, D., HEATH, D., AND 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 *Creative Evolutionary Systems*, P. J. Bentley and D. W. Corne, Eds. Morgan Kaufmann Publishers, Chapter 13, 339–365.

SECRETAN, J., BEATO, N., DAMBROSIO, D. B., RODRIGUEZ, A., CAMPBELL, A., FOLSOM-KOVARIK, J. T., AND STANLEY, K. O. 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. AND LATHAM, W. 1992. *Evolutionary art and computers*. Academic Press.

TSOUMAKAS, G. AND KATAKIS, I. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining 3,* 3, 1–13.

VEALE, T. AND HAO, Y. 2007. Comprehending and generating apt metaphors: A web-driven, case-based approach to figurative language. *AAAI Proceedings of the 22nd national conference on Artificial Intelligence 2*.

VEALE, T. AND LI, G. 1991. Creative introspection and knowledge acquisition: Learning about the world through introspective questions and exploratory metaphors. *Proceedings of the 25th AAAI Conference on Artificial Intelligence*.

WANG, W.-N. AND 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., 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.