Autonomously Communicating Conceptual Knowledge Through Visual Art

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

Abstract

In visual art, the communication of meaning or intent is an important part of eliciting an aesthetic experience in the viewer. We present a computer system, called DARCI, that is designed to automatically create original images that convey meaning. Building on previous work, we present three new components of DARCI that enhances its ability to communicate concepts through the images it creates. The first component is a model of semantic memory based on word associations that helps to provide meaning to concepts. The second component composes universal icons into a single image and then renders the image to match an associated adjective. The third component is a similarity metric that keeps the icons recognizable, but still allows for artistic elements to be discovered during the adjective rendering phase. We use an online survey to show that the system is successful at creating images that communicate concepts to human viewers.

Introduction

DARCI, Digital ARtist Communicating Intention, is a system we have built to generate original images that convey meaning. The system is part of ongoing research in the subfield of computational creativity, and is inspired by other artistic image generating systems such as Harold Cohen's AARON (McCorduck 1991) and Simon Colton's Painting Fool (Colton 2011).

Central to the design philosophy of DARCI is the notion that the communication of meaning in art is a necessary part of eliciting an aesthetic experience in the viewer (Csíkzentmihályi and Robinson 1990). DARCI is unique from other computationally creative systems in that DARCI creates images that explicitly express a given concept. Prior to this work, DARCI has been confined to only expressing adjectives through the use of global image filters (Norton, Heath, and Ventura 2010; 2011). DARCI uses a genetic algorithm to learn the image filters and parameters necessary to render a pre-existing *source image* so that it will convey specified adjectives in an interesting way. Often, due to excessive filtering and extreme parameters, this leaves the source image unrecognizable.

In this paper we introduce new capabilities to DARCI; primarily, the ability to produce original source images rather than relying upon human provided source images. DARCI composes these original source images from existing elements in order to express a wide range of concepts beyond strictly adjectives. Furthermore, in order to preserve the content of source images after applying image filters, we introduce a variation on the system's traditional image rendering technique. By polling online volunteers, we show that with these additions, DARCI is capable of creating images that convey selected concepts while maintaining the aesthetics achieved with filters.

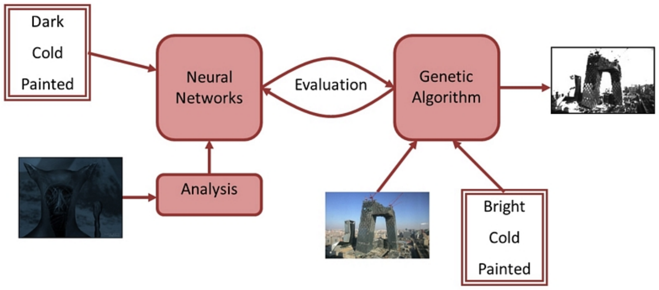


Figure 1: A diagram outlining the components of DARCI. Neural networks are trained to annotate images with adjectives. A genetic algorithm, governed by adjective annotation, is used to evolve renderings of a source image to convey specified adjectives.

Background

DARCI can be divided into two major components, the *image analysis* component, and the *image generation* component. The image analysis component learns how to annotate images with adjectives by training a series of neural networks with example images. The image generation component renders a source image so that it will visually convey an adjective using a genetic algorithm governed by the analysis component. Figure 1 outlines these two components and their interaction. In this paper, we will introduce an element of the generation component that composes a source image to match any concept (adjective or otherwise) prior to rendering it.

Image Analysis

The image analysis component uses global image features to identify general characteristics of images such as various aesthetic qualities, style, emotional impact, medium, etc. A broad range of such global features have been proven in the works of Gever, Li, Datta, Wang, and Zujovic (Gevers and Smeulders 2000; Li and Chen 2009; Datta et al. 2006; Wang, Yu, and Jiang 2006; Zujovic, Gandy, and Friedman 2007). King et al. have developed a library of these features for public use called DISCOVIR (King 2002). We use this library and one additional feature that we have designed for counting the number of highly representative hues (quantized) present in the image. In total, we use 102 features for image analysis.

To annotate images with adjectives, we train a machine learner using human labeled training data. To facilitate the acquisition of this data, we have published a website where anonymous users can label data (http://darci.cs.byu.edu). Since words can have multiple senses or meanings, we use WordNet synsets to disambiguate labels. WordNet is an ontology of words and their semantic relationships commonly used in language research (Fellbaum 1998). A *synset* is a synonym set, or a set of words that share the same meaning. A word can belong to many synsets and each synset can have many word forms. In this paper, when we refer to adjectives, we are actually referring to unambiguous WordNet synsets.

Learning to annotate images with adjectives is a *multi-label classification* problem (Tsoumakas and Katakis 2007), meaning each image can be associated with more than one adjective. To handle multi-label classification, we use a collection of artificial neural networks (ANNs) that take standardized versions of the 102 global image features as input. In reference to Colton's model of creativity (Colton 2008), we call these neural networks appreciation networks. There is an appreciation network for each adjective that has a sufficient amount of training data. As the system incrementally accumulates more data, new neural networks can be dynamically added to the collection to accommodate the new adjectives. 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 adjective.

Because a large amount of training data is needed to effectively train neural networks, and because there is a scarcity of adjective-labeled images available, we employ various strategies to augment the training data we collect. Negative training data, in particular, is difficult to come by since it is not natural for people to label images with what is *not* associated with the image, and we can't assume implicit negativity. Unfortunately, negative data is also very important for effectively training neural nets. To obtain negative data from positive data, we use the WordNet antonymy relationship. For example, if an image is labeled "hot", we assume that the image is not "cold". Another strategy we employ to increase both positive and negative data points, is to predict additional labels using correlation data that we collect (Norton, Heath, and Ventura 2010).

Image Generation

DARCI uses an evolutionary mechanism, similar to a traditional genetic algorithm, to explore the space of image filters that will render any source image according to specified adjectives while preserving specified nouns in the source image. To do this, we create a population of genotypes that evolve over many generations. Each genotype is a list of Photoshop-like filters (and their accompanying parameters) for processing the source image. Each processed source image is called a phenotype. Figure 2 gives an example of a genotype and its phenotype.

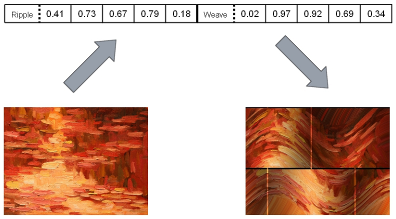


Figure 2: Sample genotype (top) applied to a source image (left) resulting in the phenotype (right). The genotype is a list of image filters with parameters. "Ripple" and "Weave" are the names of two (of ninety-two) possible filters.

Every generation, each phenotype is evaluated by analyzing it with the appreciation networks corresponding to the specific adjectives required. The appreciation networks provide a score indicating the strength of a match found in the phenotype. The most fit, or highest scoring, phenotypes' corresponding genotypes are preserved for the next generation of the algorithm. New genotypes are created, to replace low scoring genotypes, by combining filters between the most fit genotypes in a process known as *crossover*. Finally, a small fraction of genotypes are *mutated*: filter settings are slightly shifted.

This process is repeated until either a genotype emerges with a score above some threshold, or a specified amount of time elapses. The highest scoring phenotypes are returned as successful artifacts. More details into this process can be found in a previously published paper (Norton, Heath, and Ventura 2011).

Methodology

In this section, we introduce several new advances to DARCI that enhance the system's capability to communicate intended meaning in an aesthetic fashion: a semantic memory model enables the system to express any concept with associations that are relatable by a human audience, an image composer allows the system to compose concrete representations of the concepts learned by the semantic memory model into source images for DARCI to render, and finally a new metric governing the evolution of new images enables the system to more effectively convey these concepts in final renderings. We also describe an online survey that we use to evaluate the success of these additions.

Semantic Memory Model

In cognitive psychology, the term *semantic memory* means the memory of meaning and other concept-based knowledge, and allows people to consciously recall general information about the world. We argue that in order to be creative, there needs to be intent and purpose behind what is being created. How can a system intentionally create an image about 'war' if the system has no knowledge about what 'war' means? In order for DARCI to visually communicate a more advanced concept, DARCI needs to have some internal knowledge (or understanding) of that concept (i.e, its own semantic memory).

This question of what gives words (or concepts) meaning has been debated for years; however, it is commonly agreed that a word, at least in part, is given meaning by how the word is used in conjunction with other words (i.e., its context) (Erk 2010). Many computational models of semantic memory consist of building associations between words (Sun 2008; De Deyne and Storms 2008). These word associations essentially form a large graph that is typically referred to as a *semantic network*. For example, if the system wanted to communicate the concept 'dog', other words associated with 'dog' are retrieved, words like 'fur', 'bark', 'tail', 'poodle' 'leash', 'pound', etc. These words could be attributes of 'dogs', things 'dogs' do, types of 'dogs', or other objects/places that commonly occur with 'dogs'. These associated words provide a level of meaning to the concept 'dog', which will help the system to successfully convey the concept to others.

Word associations are commonly acquired in one of two ways: from people and automatically by inferring them from a corpus. Here we describe a computational model of semantic memory that combines human free association norms with a simple corpus-based approach. The idea is to use the human word associations to capture general knowledge and then fill in the gaps using the corpus method.

Lemmatization and Stop Words In gathering word associations, we use the standard practice of removing stop words (words like 'the' and 'of') and lemmatizing (combining different forms of the same word). WordNet maintains a database of word forms and hence, we use WordNet to perform the lemmatization (Fellbaum 1998). It should be noted, however, that lemmatization with WordNet has its limits. For example, we cannot lemmatize a word across different parts of speech (noun, verb, adjective, etc). For example, 'jump' and 'jumping' will remain separate words because 'jumping' could be the gerund form of the verb 'jump' or it could be a noun (i.e., the act of 'jumping'). Since the part of speech is not provided for individual words, we must account for all parts of speech, hence words like 'relax', 'relaxing' and 'relaxation' remain separate words.

Free Association Norms One of the most common means of gathering word associations from people is through *Free Association Norms* (FANs), which is done by asking hundreds of human volunteers to provide the first word that comes to mind when given a cue word. This technique is able to capture many different types of word associations including word co-ordination (pepper, salt), collocation (trash, can), super-ordination (insect, butterfly), synonymy (starving, hungry), and antonymy (good, bad). The association strength between two words is simply a count of the number of volunteers that said the second word given the first word. FANs are considered to be one of the best methods for understanding how people, in general, associate words in their own minds (Nelson, McEvoy, and Schreiber 1998). In our model we use two preexisting databases of FANs: The Edinburgh Associative Thesaurus (Kiss et al. 1973) and University of Florida's Word Association Norms (Nelson, McEvoy, and Schreiber 1998).

It should be noted that in this model we consider word associations to be undirected. In other words, if word A is associated with word B, then word B is associated with word A. Hence, when we encounter data in which word A is a cue for word B and word B is also a cue for word A, we combine them into a single association pair by adding their respective association strengths. Between these two databases, there are a total of 19,327 unique words and 288,069 unique associations. From now on, we will refer to these associations as *human data*.

Corpus Inferred Associations Discovering word associations from a corpus is typically accomplished using methods from a family of techniques called *Vector Space Models* (Turney and Pantel 2010), which uses a matrix for keeping track of word counts either co-occurring with other words (term × term matrix) or within each document (term × document matrix).

One of the most popular vector space models is *Latent Semantic Analysis* (LSA) (Deerwester et al. 1990). LSA is based on the idea that similar words will appear in similar documents (or contexts). LSA builds a term × document matrix from a corpus and then performs a technique called Singular Value Decomposition (SVD), which essentially reduces the large sparse matrix to a low-rank approximation of that matrix along with a set of vectors, each representing a word (as well as a set of vectors for each document). These vectors also represent points in semantic space, and the closer words are to each other in this space, the closer they are in meaning (and the stronger the association between words).

Another popular method is the *Hyperspace Analog to Language* (HAL) model (Lund and Burgess 1996). This model is based on the same idea as LSA, except the notion of context is reduced more locally to a word co-occurrence window of ±10 words instead of an entire document. Thus, the HAL model builds a term × term matrix of word co-occurrence counts from a corpus. HAL then uses the co-occurrence counts directly as vectors representing each word in semantic space. The term × term matrix is advantageous because the size of the matrix is invariant to the size of the corpus, it is also argued by some that it is more congruent to human cognition than the term × document matrix used in LSA (Wandmacher, Ovchinnikova, and Alexandrov 2008; Burgess 1998).

The corpus component of our model is constructed similarly to HAL but with some important differences. We restrict the model to the same number of unique words as the human-generated free associations, building a 19,327 × 19,327 (term × term) co-occurrence matrix M using a co-occurrence window of ±50. To account for the fact that common words will have generally higher co-occurrence counts, we scale these counts by weighting each element of the matrix by the inverse of the total frequency of both words at each element. This is done by considering each element $M_{i,j}$, then adding the total number of occurrences of each word (i and j), subtracting out the value at $M_{i,j}$ (to avoid counting it twice), then dividing $M_{i,j}$ by this computed number, as follows:

$$M_{i,j} \leftarrow \frac{M_{i,j}}{(\sum_{i} M_{i,j} + \sum_{j} M_{i,j} - M_{i,j})} \tag{1}$$

The result could be a very small number and hence, we then also normalize the values between 0 and 1.

For our corpus we use Wikipedia, as it is large, easily accessible, and covers a wide range of human knowledge (Denoyer and Gallinari 2006). Once the co-occurrence matrix is built from the entire text of Wikipedia, we use the weighted/normalized co-occurrence values themselves as association strengths between words. This approach works, since we only care about the strongest associations between words, and it allows us to reduce the number of irrelevant associations by ignoring any word pairs with a co-occurrence count less than some threshold. We chose a threshold of 100 (before weighting), which provides a good balance of producing a sufficient number of associations, while reducing the number of irrelevant associations. When looking up a particular word, we return the top n other words with the highest weighted/normalized co-occurrence values. This method, which we will call *corpus data* from now on, gives a total of 4,908,352 unique associations.

Combining Word Associations Since each source (human and corpus) provide different types of word associations, a combination of these methods into a single model has the potential to take advantage of the strengths of each method. The hypothesis is that the combined model will better communicate meaning to a person than either model individually because it presents a wider range of associations.

Our method merges the two separate databases into a single database before querying it for associations. This method assumes that the human data contains more valuable word associations than the corpus data because the human data is typically used as the gold standard in the literature. However, the corpus data does contain some valuable associations not present in the human data. The idea is to add the top n associations for each word from the corpus data to the human data but to weight the association strength low. This is beneficial for two reasons. First, if there are any associations that overlap, adding them again will strengthen the association in the combined database. Second, new associations not present in the human data will be added to the combined database and provide a greater variety of word associations. We keep the association strength low because we want the corpus data to reinforce, but not dominate, the human data.

To do this, we first copy all word associations from the human data to the combined database. Next, let W be the set of all 19,327 unique words, let $A_{i,n} \subseteq W$ be the set of the top n words associated with word $i \in W$ from the corpus data, let $score_{i,j}$ be the association strength between words i and j from the corpus data, let max_i be the maximum association score present in the human data for word i, and let θ be a weight parameter. Now for each $i \in W$ and for each $j \in A_{i,n}$, the new association score between words i and j is computed as follows:

$$score_{i,j} \leftarrow (max_i \cdot \theta) \cdot score_{i,j} \tag{2}$$

This equation scales $score_{i,j}$ (which is already normalized) to lie between 0 and a certain percentage (θ) of max_i. The n associated words from the corpus are then added to the combined database with the updated scores. If the word pair is already in the database, then the updated score is added to the score already present. For the results presented in this paper we use $n = 20$ and $\theta = 0.2$, which were determined based on preliminary experiments. After the merge, the combined database contains 443,609 associations.

Image Composer

The semantic memory model effectively uses word associations to break down a concept into simpler concepts that together represent the whole. If a concept is simple enough, it can be represented visually with a single icon. For example, the concept 'rock' can be visually represented with a picture of a 'rock'. The idea is to gather a set of icons that together represent the overall concept and compose those icons into a single image. The image is then given to the adjective rendering component of DARCI which renders the image to match some adjective associated with the concept.

We use a collection of icons provided by *The Noun Project*, whose goal is to build a repository of symbols/icons that can be used as a visual language (nou 2013). The icons are intended to be simple visual representations of any noun and are published by various artists under the Creative Commons license. Currently, The Noun Project provides 6,334 icons (each 420 × 420 pixels) representing 2,535 unique nouns and is constantly growing.

When given a concept, DARCI first uses the semantic memory model to retrieve all words associated with the given concept including itself. These word associations are filtered by returning only nouns that DARCI has icons for and adjectives that DARCI has appreciation networks for. The nouns are sorted by association strength and no more than the top 15 are kept. For each noun, multiple icons are usually available and one or two of these icons are are chosen at random to create a set of icons for use in composing the image.

The icons in the set are scaled to between 25% and 100% of their original size according to their association strength rank. Let I be the set of icons, and let $r : I \rightarrow [0, |I| - 1]$ be the rank of icon $i \in I$, where the icon with rank 0 corresponds to the noun with the highest association strength. Finally, let ϕ_i be the scaling factor for icon i, which is computed as follows:

$$\phi_i \leftarrow 1 - \frac{0.75}{|I|} \cdot r(i) \tag{3}$$

An initial blank white image of size 2000 × 2000 pixels is created and the set of scaled icons are drawn onto the blank image at random locations. The only constraints being that no icons are allowed to overlap and no icons are allowed to go off the edge of the image. The result is a collage of icons that represents the original concept. DARCI then randomly selects an adjective from the set returned by the semantic memory model weighted by each adjective's association strength. DARCI uses its adjective rendering component to render the collage image, now a source image, according to the selected adjective. The final image will both be artistic and in some way communicate the concept to the viewer.

Similarity Metric

In previous papers, the fitness function DARCI used to evaluate artifacts included two components, the *adjective metric* and the *interest metric*. The adjective metric is simply the output of the neural network trained on the adjective of interest. Using this metric alone, the source image is usually completely obliterated after only a few generations of evolution. This is because the neural networks are trained entirely on the global features of a variety of images. A high output from these networks indicates an artifact that has strong association with these generalized features without any consideration given to the source image.

The interest metric is a measure of how "interesting" the image is with respect to the source image. Is is based on the subjective assumption that an interesting artifact would be one that is different from the source while still maintaining some semblance of it. The metric is a function of the number of global features that are similar (within some threshold) between the source image and the artifact, and attributes a high score to those images that fall within some middle range of similar features (Norton, Heath, and Ventura 2011). The interest metric was introduced in order to attempt to alleviate the problem of completely loosing the source image, while simultaneously not rewarding an artifact that looks *too* similar to the source. In previous research, we have normalized these two metrics to have an equal weight on the fitness function. While an improvement, features of the source image are still often eliminated after many generations of evolution using this combined metric. We hypothesize that this is because the interest metric is only based on global features that are not specific enough to encapsulate much of what defines an image.

In this paper we use a similarity metric that borrows from the growing research of bag-of-visual-word models (Csurka et al. 2004; Sivic et al. 2005; Kandasamy and Rodrigo 2010) to analyze local features, rather than global ones. Visual words are quantized local image features that takes their cue from natural language processing. A dictionary of visual words are defined for any given scope by extracting local interest points from a large number of representative images, and then clustering them (typically with k-means) by their features into n clusters, where n is the desired dictionary size. With this dictionary, visual words can be extracted from images by determining which clusters the images' local interest points belong to. A bag-of-visual-words can be created by organizing the visual word counts for a given image into a fixed vector. This model is analogous to the bag-of-words for documents in natural language processing.

For our similarity metric, which we call the *local similarity metric*, we first create a bag-of-visual-words for the source image and the artifact, and then calculate the euclidean distance between these two vectors. This metric has the effect of measuring the quantity of interest points that coincide between the two images. We hypothesize that using this locally-based metric in the fitness function will allow DARCI to produce images that better maintain the source images' structure.

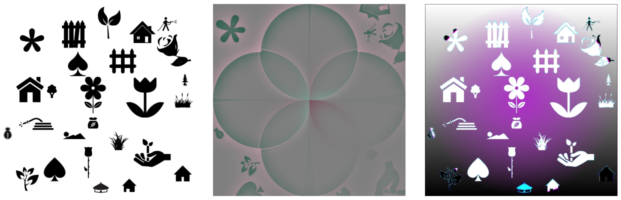
In this paper, we use the standard SURF detector and descriptor to extract interest points and their features from images (Herbert Bay 2008). We build the visual word dictionary by extracting SURF interest points from a database of universal icons obtained at The Noun Project (nou 2013). At the time of this paper we have extracted 6334 icons, which results in more than two hundred thousand interest points. These are then clustered into 1000 visual words using Elkan k-means (Elkan 2003). Once the euclidean distance, d, between the source image's and the artifact's bags-of-visual-words is calculated, the metric, S, is calculated to provide a value between 0 and 1 as follows: $S = MAX(\frac{d}{100}, 1)$, where the constant 100 was chosen through preliminary observation.

Online Survey

With DARCI, we are interested in a system that can create images that both communicate meaning and are aesthetically interesting. For this paper, we have developed a survey to test our most recent attempts at conveying concepts while rendering images that are perceived as creative.

The survey asks users to evaluate collages generated for ten concepts across three rendering techniques. The ten concepts were chosen to cover a variety of topics including abstract ideas and concrete objects. The abstract concepts selected were 'adventure', 'love', 'music', 'religion', and 'war'. The concrete concepts were 'bear', 'cheese', 'computer', 'fire', and 'garden'.

We refer to the three rendering techniques as *unrendered*, *traditional*, and *advanced*. For *unrendered*, no rendering is applied—these are the plain collages. For the other two techniques, the images are rendered as described above using one of two fitness functions to govern the evolutionary mechanism. With *traditional*, the fitness function is the same as we have used in previous research, the adjective and interest metrics are each given a weight of 0.5. With *advanced* on the other hand, we introduce the new local similarity metric. Here the adjective metric is given a weight of 0.5, while the interest and local similarity metrics are each given a weight of 0.25. For each rendering technique and image, DARCI returned the 40 highest ranking images discovered over the period of around 90 generations. We then selected from each concept and technique, the image that we felt best conveyed the intended concept while appearing aesthetically interesting. An example image that we selected from each rendering technique can be seen in Figure 3.



(a) unrendered (b) traditional (c) advanced

Figure 3: Example images for the three rendering techniques representing the concept ‘garden’. The original icons used in these particular images are from various artists at The Noun Project (Zubin et al. 2013).

To query the users about each image, we follow the survey template that we developed previously to study the perceived creativity of images rendered with different adjectives (Norton, Heath, and Ventura 2013). In this study, we presented users with six five-point Likert items (Likert 1932) per image; volunteers were asked how strongly they agreed or disagreed (on a five point scale) with each statement as it pertained to one of DARCI’s images. The six statements we used were (abbreviation of item in parentheses):

I like the image. (*like*)

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

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

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

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

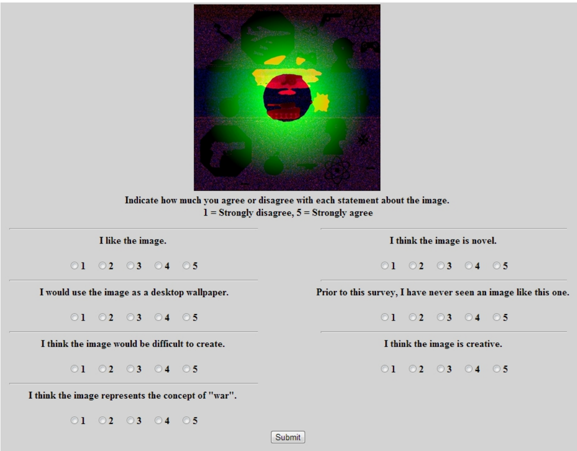
I think the image is creative. (*creative*)

In (Norton, Heath, and Ventura 2013) we showed that the first five statements correlated strongly with the sixth, “I think the image is creative”, justifying this test as an accurate evaluation of an image’s subjective creativity. In this paper, we use the same six Likert items and add a seventh to determine how effective the images are at conveying their intended concept. The seventh statement we include is:

I think the image represents the concept of “____.” (*concept*)

where the blank space contains the intended concept. Figure 4 shows the Likert items with an accompanying image as they appear in the survey.

To avoid fatigue, volunteers were only presented with images from one of the three rendering techniques mentioned previously. The technique was chosen randomly and then the images were presented to the user in a random order. To help gauge the results, three dummy images were introduced into the survey for each technique. These dummy images were selected from the results of preliminary experimentation and assigned an arbitrary concept for the survey. Unfiltered dummy collages were added to the unrendered set of images, while rendered versions were added to the traditional and advanced sets of images. The three dummy concepts were: ‘restaurant’, ‘water’, and ‘freedom’. An example unrendered dummy image for the concept of ‘freedom’ is shown in Figure 5. In total, each volunteer was presented with 13 images.



Indicate how much you agree or disagree with each statement about the image.
1 = Strongly disagree, 5 = Strongly agree

I like the image.

○1 ○2 ○3 ○4 ○5

I think the image is novel.

○1 ○2 ○3 ○4 ○5

I would use the image as a desktop wallpaper.

○1 ○2 ○3 ○4 ○5

Prior to this survey, I have never seen an image like this one.

○1 ○2 ○3 ○4 ○5

I think the image would be difficult to create.

○1 ○2 ○3 ○4 ○5

I think the image is creative.

○1 ○2 ○3 ○4 ○5

I think the image represents the concept of "war".

○1 ○2 ○3 ○4 ○5

Figure 4: A sample image from the survey, in this case the image is the concept ‘war’ rendered with the ‘advanced’ technique. Below the image, the seven five-point Likert items are presented.

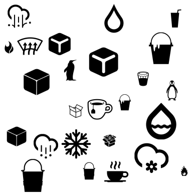


Figure 5: An example of one dummy image presented to users to gauge survey results. This image is the unrendered image for the concept of ‘freedom’.

Results

A total of 119 anonymous individuals participated in the online survey. Each person evaluated an average of 9 images and each image was evaluated by an average of 27 people for a total of 1069 data points.

The three dummy images for each rendering technique are used as a baseline for the concept statement. The results of the dummy images versus the valid images are show in Figure 6. The average concept rating for the valid images is significantly better than the dummy images, which shows that the intended meaning is successfully conveyed to human viewers more reliably than an arbitrary image. These results confirm that the intelligent use of icons is beneficial for the visual communication of meaning. The ratings for the other statements are also generally lower for the dummy images than for the valid images. It seems as though the ability of a visual artist to express meaning to a viewer is an important factor in attributing creativity to the artist. The primary difference between the rendering of the dummy images versus the valid images is that the dummy images were created for a different concept than the one they were attributed to in the survey. Having the concept not match the image seems to negatively influence how the users rate the other statements in the survey. This provides some level of

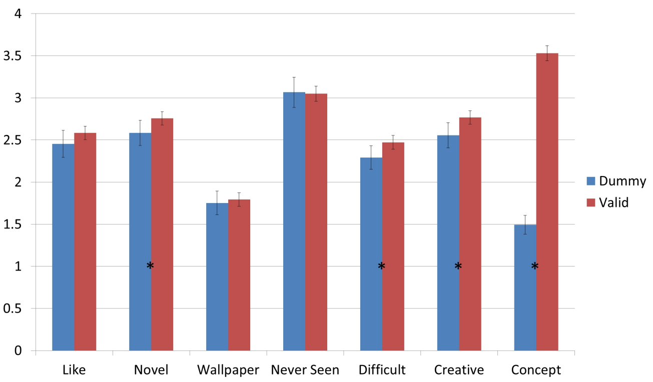


Figure 6: The average rating from the online survey for all seven statements comparing the dummy images with the valid images. The valid images were more successful at conveying the intended concept than the dummy images by a significant margin. Results marked with an asterix (*) indicate statistical significance using the two tailed independent t-test. The lines at the top of each bar show the 95% confidence interval for each value.

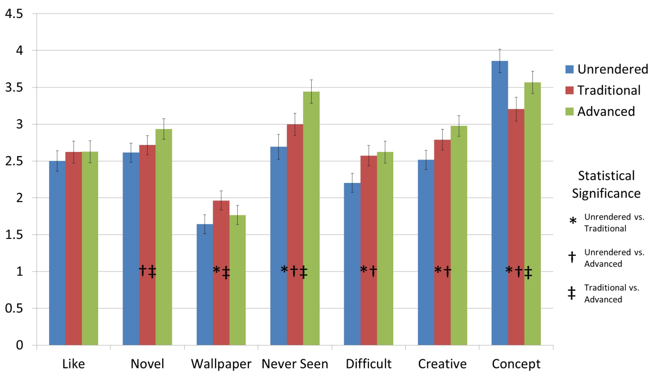


Figure 7: The average rating from the online survey for all seven statements comparing the three rendering techniques. The unrendered technique is most successful at representing the concept, while the advanced technique is generally considered more novel and creative. Statistical significance was calculated using the two tailed independent t-test. The lines at the top of each bar show the 95% confidence interval for each value.

evidence to support the notion that the perceived purpose (or intent) behind creating artifacts is a factor in attributing creativity to a system.

The results of the three rendering techniques (unrendered, traditional, and advanced) for all seven statements are shown in Figure 7. The unrendered images are generally the most successful at communicating the intended concepts. This is likely because the objects/icons in the unrendered images are left undisturbed and are therefore more clear and discernible. The rendered images (traditional and advanced) often distort the icons in a way that makes them less cohesive and less discernible and can thus take away from the intended meaning. However, the trade-off is that the unrendered images are generally considered less likable, less novel, and less creative than the rendered images. The advanced images are generally considered more novel and creative than the traditional images, but the traditional images are liked slightly more. The advanced images also convey the intended meaning more reliably than the traditional images, which indicates that the similarity metric is successfully finding a better balance between adding artistic elements and allowing the icons/objects to still be recognizable.

The results comparing the abstract concepts with the concrete concepts are show in Figure 8. For all seven statements, the abstract concepts are, on average, rated higher than the concrete concepts. One possible reason for this is that concrete concepts are not easily broken down to a collection of other concrete objects because they can already be represented as a single icon. The nouns returned by the semantic memory model are usually other related objects, but it then becomes difficult to tell which object is the concept in question. For example, the concept 'bear' returns nouns like 'cave', 'lion', 'forest', and 'wolf', which are all related, but don't provide much indication that the intended concept is 'bear'. A person might be more inclined to think the concept is more general, such as 'wildlife'. Another possible reason why abstract concepts perform better than concrete concepts is because abstract concepts allow a wider range of interpretation and are generally more interesting. For example, the concept 'cheese' would seem to be pretty straightforward to most people, while the concept 'love' could have variable meaning to different people in different circumstances. Hence, the resulting images are generally considered more likable, more novel, and more creative than the concrete images.

Conclusions and Future Work

We have presented three new components of a computer system, DARCI, capable of communicating specified concepts through the images it creates. The first new component is a model of semantic memory that provides the system with a level of meaning for concepts through word associations. The second component uses the word associations from the semantic memory model to retrieve universal icons and compose them into a single image, which is then rendered in the manner of an associated adjective. The third component is a new similarity metric used during the adjective rendering phase that preserves the discernibility of the icons, but still allows for artistic elements to be discovered.

We used an online survey to evaluate the system and show that DARCI is significantly better at expressing the meaning of concepts through the images it creates than an arbitrary image. We show that the new similarity metric allows DARCI to find a better balance between adding interesting artistic qualities and keeping the icons/objects recognizable. We show that using word associations and universal icons in an intelligent way is beneficial for conveying meaning to human viewers. Finally, we show that there is some degree of

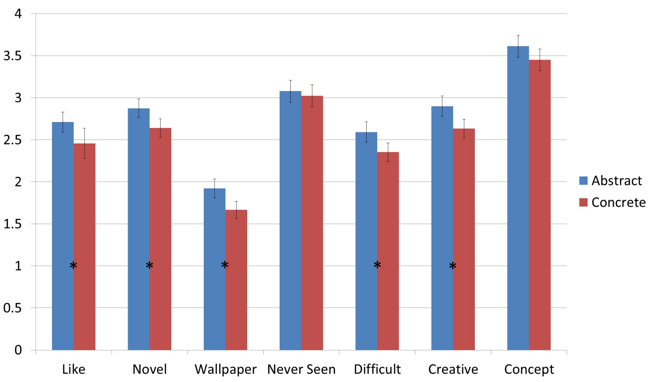


Figure 8: The average rating from the online survey for all seven statements comparing the abstract concepts with the concrete concepts. The abstract concepts generally received higher ratings for all seven statements. Results marked with an asterix (*) indicate statistical significance using the two tailed independent t-test. The lines at the top of each bar show the 95% confidence interval for each value.

correlation between how well an image communicates the intended concept and how well liked, how novel, and now creative the image is considered to be.

The semantic memory model provides DARCI with some conceptual knowledge that is necessary for determining how to compose and render an image that is unique and meaningful for each concept. We hypothesize that this is a significant component of a creative system because it could allow the system to make decisions and reason about common world knowledge. In future research we plan to do a more direct comparison of the images created by DARCI with images created by human artists and look closely at how semantic memory contributes to the creative process. We plan to improve the semantic memory model by going beyond word-to-word associations and building associations between words and other objects (such as images). This would require expanding DARCI's image analysis capability to be able to automatically detect and annotate objects present in an arbitrary image. The similarity metric presented in this paper is a step in that direction. An improved semantic memory model could also help provide DARCI the ability to discover its own topics (i.e., find its own inspiration) and learn how to compose icons together in more meaningful ways.

References

Burgess, C. 1998. From simple associations to the building blocks of language: Modeling meaning in memory with the HAL model. *Behavior Research Methods, Instruments, & Computers* 30:188–198.

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

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

Csíkzentmihályi, M., and Robinson, R. E. 1990. *The Art of Seeing*. The J. Paul Getty Trust Office of Publications.

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

De Deyne, S., and Storms, G. 2008. Word associations: Norms for 1,424 Dutch words in a continuous task. *Behavior Research Methods* 40(1):198–205.

Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; and Harshman, R. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41(6):391–407.

Denoyer, L., and Gallinari, P. 2006. The Wikipedia XML corpus. In *INEX Workshop Pre-Proceedings*, 367–372.

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

Erk, K. 2010. What is word meaning, really?: (and how can distributional models help us describe it?). In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, 17–26. Stroudsburg, PA, USA: Association for Computational Linguistics.

Fellbaum, C., ed. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

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

Herbert Bay, Andreas Ess, T. T. L. V. G. 2008. Speeded-up robust features (surf). *Computer Vision and Image Understanding* 110:346–359.

Kandasamy, K., and Rodrigo, R. 2010. Use of a visual word dictionary for topic discovery in images. In *Proceedings of 5th International Conference on Information and Automation for Sustainability*, 510–515.

King, I. 2002. Distributed content-based visual information retrieval system on peer-to-pear(p2p) network. http://appsrv.cse.cuhk.edu.hk/~miplab/discovir/.

Kiss, G. R.; Armstrong, C.; Milroy, R.; and Piper, J. 1973. An associative thesaurus of English and its computer analysis. In Aitkin, A. J.; Bailey, R. W.; and Hamilton-Smith, N., eds., *The Computer and Literary Studies*. Edinburgh, UK: University Press.

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

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

Lund, K., and Burgess, C. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence.

Behavior Research Methods, Instruments, & Computers 28:203–208.

McCorduck, P. 1991. *AARON's Code: Meta-Art, Artificial Intelligence, and the Work of Harold Cohen*. W. H. Freeman & Co.

Nelson, D. L.; McEvoy, C. L.; and Schreiber, T. A. 1998. The University of South Florida word association, rhyme, and word fragment norms. http://www.usf.edu/FreeAssociation/.

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.

Norton, D.; Heath, D.; and Ventura, D. 2013. Finding creativity in an artificial artist. *Journal of Creative Behavior, to appear*.

2013. The noun project. `http://thenounproject.com`.

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.

Sun, R. 2008. *The Cambridge Handbook of Computational Psychology*. New York, NY, USA: Cambridge University Press, 1st edition.

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

Turney, P. D., and Pantel, P. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research (JAIR)* 37:141–188.

Wandmacher, T.; Ovchinnikova, E.; and Alexandrov, T. 2008. Does latent semantic analysis reflect human associations? In *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*, 63–70.

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.

Zubin, A.; Brain, B.; Caughey, E.; Fisher, R.; Patel, P.; Barriga, R.; dsathiyaraj; Bristol, J.; Fortnum, A.; Koltringer, M.; MacKenzie, B.; Schlosman, H.; Pedrazzoli, M.; Endale, M.; Agpoon, G.; and Eckert, J. 2013. The noun project. `http://thenounproject.com`.

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