Establishing Appreciation in a Creative System

David Norton, Derral Heath, Dan Ventura

Computer Science Department
Brigham Young University
dnorton@byu.edu, dheath@byu.edu, ventura@cs.byu.edu

Abstract. Colton discusses three conditions for attributing creativity to a system: appreciation, imagination, and skill. We describe an original computer system (called DARCI) that is designed to eventually produce images through creative means. We show that DARCI has already started gaining appreciation, and has even demonstrated imagination, while skill will come later in her development.

1 Introduction

While several theoretical frameworks for creativity have been proposed, actually building a system that applies these frameworks is difficult. We are developing an original system designed to implement and integrate concepts proposed by researchers such as Boden, Wiggens, Ritchie, and Colton. Our system, DARCI (Digital ARtist Communicating Intention), will produce images that are not only perceived by humans as creative products, but that are also produced through arguably creative processes. This paper represents our work with only the first component of DARCI, that of learning about the domain of visual art. We will discuss why this is an important step in the creative process in terms of Colton's creative tripod concept [3], describe how DARCI is learning about this domain, and finally demonstrate DARCI's current level of development.

Colton discusses three attributes that must be perceived in a system to consider it creative: appreciation, imagination, and skill. In order for DARCI to be appreciative of art, she needs to first acquire some basic understanding of art [3]. For example, in order for DARCI to appreciate an image that is gloomy, she has to first recognize that it is gloomy. To facilitate this, we are teaching DARCI to associate low-level image features with artistic descriptions of the image. Currently, DARCI has learned how to associate 150 different descriptors to images. Furthermore, she can essentially interpret an image by selecting a specific combination of these descriptors for the image in question, thus demonstrating a degree of imagination. This will also facilitate communication with DARCI's audience, enhancing the perception of appreciation and imagination. DARCI cannot yet produce any images and so does not yet demonstrate skill in the sense that Colton prescribes. However, at the end of this paper we will show how DARCI's understanding of the art domain will be instrumental to her production of original images.

2 Image Feature Extraction

Before DARCI can form associations between image features and descriptive words, the appropriate image features for the task must be selected. These need to be low-level features that characterize the various ways that an image can be appreciated.

There has been a large amount of research done in the area of image feature extraction. King and Gevers deal with Content Based Image Retrieval (CBIR) [2][6]. CBIR relies heavily on extracting image features which can then be compared and used when searching for images with specific content. CBIR systems look at characteristics such as an image's color, light, texture, and shape. Datta and Li propose several image features that look at these same characteristics to assess the aesthetic quality of images [4][7].

Wang deals with image retrieval specific to emotional semantics [10][9]. The goal is to search for images that have specific emotional qualities such as happy, gloomy, showy, etc. Zujovic tries to classify a painting into one of six different genres: Abstract, Expressionism, Cubism, Impressionism, Pop Art and Realism [11]. All of these researchers have proposed image features that focus on color, light, texture, and shape. Of these image features, we have selected 102 of the more common ones to use in DARCI. As with prior research, our set of image features is broken down into characteristics relating to color, light, texture and shape.

Color and light play a significant role in the emotion and meaning conveyed in images. Colors have often been associated directly with emotions. For example, red can mean anger and frustration while blue can mean sad and depressed. Likewise with light, a dark image could mean gloomy or scary while a bright image could denote happiness or enthusiasm. Texture and shape features also play a significant role in the meaning and emotion of an image. For example, a cluttered and busy image could indicate feelings of anxiety or confusion. An image that is blocky and structured could indicate feelings of stability and security. We extract eight color features, four light features, 50 texture features and 40 shape features as follows:

Color & Light:

1. Average Red, Green, and Blue
2. Average Hue, Saturation, and Intensity
3. Unique Hue count (20 buckets)
4. Average Hue, Saturation, and intensity contrast
5. Dominate hue
6. Percent of image that is the dominate hue

Shape:

1. Geometric Moment
2. Eccentricity
3. Invariant Moment (5x vector)
4. Legendre Moment
5. Zernike Moment
6. Psuedo-Zernike Moment
7. Edge Direction Histogram (30 bins)

Texture:

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

It is not the purpose of this paper to go into detail about the image features we extracted. These features were selected based on the results of the research previously mentioned.

3 Visuo-Linguistic Association

DARCI forms an appreciation of art by making associations between image features and descriptions of the images. An image can be described and appreciated in many ways: by the subject of the image, by the aesthetic qualities of the image, by the emotions that the image evokes, by associations that can be made with the image, by the meanings found within the image, and possibly others. To teach DARCI how to make associations with such descriptors, we present her with images labeled appropriately. Ideally we would like DARCI to understand images from all of these perspectives. However, because the space of all possible images and their possible descriptive labels is enormous, we have taken measures to reduce the descriptive label space to one that is tractable. Specifically, we have reduced descriptive labels exclusively to delineated lists of adjectives.

3.1 WordNet

We use WordNet's [5] database of adjectives to give us a large, yet finite, set of descriptive labels. Even though our potential labels are restricted, the complete set of WordNet adjectives can allow for images to be described by their emotional effects, most of their aesthetic qualities, many of their possible associations and meanings, and even, to some extent, by their subject.

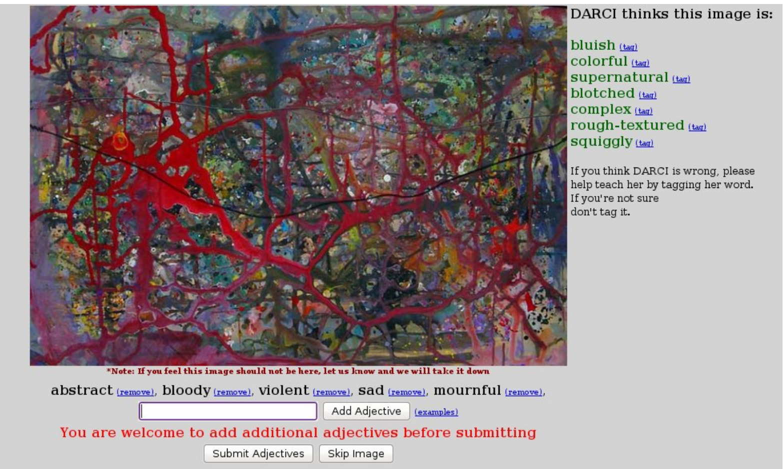
In WordNet, each word belongs to a synset of one or more words that share the same meaning. If a word has multiple meanings, then it can be found in multiple synsets. For example, the word "dark" has eleven meanings, or senses, as an adjective. Each of these senses belongs to a unique synset. The synset for the sense of "dark" that means "stemming from evil characteristics or forces; wicked or dishonorable", also contains senses of the words "black" and "sinister". Our image classification labels actually consist of a unique synset identifier, rather than the adjectives themselves.

3.2 Learning Method

In order to make the association between image features and descriptors, we use a series of artificial neural networks trained incrementally with backpropagation. A training instance is defined as the image features for a particular image paired with a single synset label. We create a distinct neural network, with a single output node, for each synset that has a sufficient amount of training data. For the results presented in this paper, that threshold is eight training instances. Enforcing this threshold ensures a minimum amount of training data for each synset. As we incrementally accumulate data, more and more neural networks are created to accommodate the new synsets that pass the threshold. This process ensures that neural networks are not created for synsets that are either too obscure or occur only accidentally. Shen, *et al.* employ a similar approach for handling non-mutually exclusive labels to good effect using SVMs instead of ANNs [8].

4 Obtaining Data Instances

To collect training data, we have created a public website for training DARCI [1]. From this website, users are presented with a random image and asked to provide adjectives



DARCI thinks this image is:

bluish (tag)
colorful (tag)
supernatural (tag)
blotched (tag)
complex (tag)
rough-textured (tag)
squiggly (tag)

If you think DARCI is wrong, please help teach her by tagging her word. If you're not sure don't tag it.

*Note: If you feel this image should not be here, let us know and we will take it down

abstract (remove), bloody (remove), violent (remove), sad (remove), mournful (remove),

You are welcome to add additional adjectives before submitting

Fig. 1. Screenshot of the website used to train DARCI. Below the image are adjectives that the user has entered as well as a text box for entering new adjectives. To the right of the image are seven adjectives that DARCI has attributed to the image. Image courtesy of Mark Russell.

that describe the image (Figure 1). 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. We keep track of the results in an SQL database from which we can train the appropriate neural networks. As of this writing, we have obtained close to 6000 data points this way. While this is still only a small fraction of the amount of data we will need, it has proven satisfactory for some adjectives as we will show.

While there are 18,156 adjective synsets in WordNet, it is not necessary for DARCI to learn all of them. In the set of roughly 6,000 data instances we have obtained so far, only 1,176 unique synsets have occurred. Of those unique synsets, almost half have only a single example. There will be many synsets that will never meet our threshold of eight instances, thus making the association task more manageable.

The total number of synsets that have at least eight data points (our threshold for creating a neural net) is currently 150. This means that DARCI essentially "knows" 150 synsets at the writing of this paper. Keep in mind that many of those synsets contain several senses, so the number of adjectives DARCI effectively "knows" is actually much higher. As DARCI is currently nascent, this number will continue to grow.

4.1 Amplifying Data

We have been faced with two fundamental problems with regards to training data. First, all of the training data that we have examined so far is exclusively positive training data (i.e. the training data only indicates what an image *is*, not what it *is not*). It is very difficult to train ANNs without negative examples as well. The second problem is a paucity of training instances. ANNs require a lot of training data to converge and currently, of the 150 synsets known to DARCI, there are on average just over twenty three positive data instances per synset.

We have employed two methods for obtaining negative data. The first method utilizes the antonym attribute of adjectives in WordNet. Anytime an image is labeled with an adjective, we create a negative data point for all antonyms of that adjective. Second, on DARCI's website, we allow users to directly create negative examples for adjectives that DARCI knows. For each image presented to the user, DARCI lists seven adjectives that she associates with the image (Figure 1). The user is then allowed to flag those labels that are not accurate. This creates strictly negative examples. This method also allows DARCI to demonstrate to the user her current interpretation of an image. Using these methods, we have built up more negative data points than positive ones.

In order to help compensate for shortages in training data, for each new data instance that is presented to DARCI, a variable number of old data instances belonging to the same synset, are reintroduced to the neural net in question. In addition to reintroducing old material, a variable number of prior data instances that do *not* belong to the same synset, but that are statistically correlated, are introduced to the neural net in question. These guessed data instances provide DARCI with more data for each synset than she is in fact receiving, and allow DARCI to take advantage of correlations in labels that are lost by using unique neural nets for each synset. We perform these data expansion strategies to both the positive and negative data instances and do so in a manner that attempts to balance the amount of negative and positive data that DARCI receives for each synset.

The combination of adding negative data instances, recycling old data instances, guessing correlations with other synsets, and using these guesses to balance positive and negative training instances, greatly amplifies the amount of training data presented to DARCI.

5 Interpreting Images

When presented with an image, DARCI takes the output of each synset's neural net given the image features, and treats that output as a score. But DARCI currently knows 150 synsets, so how does she choose which of the synsets to label the image with? The easiest solution would be to either take all synsets with a score above a specific threshold, or take the top n synsets. However, despite our attempts to amplify the data, some synsets continue to be lacking in training instances. The neural networks for these synsets should not be given as much weight in determining the relevance of an adjective for a particular image. Thus, we use Equation 1 for modifying each neural network's output value to create a new score that takes DARCI's confidence about a particular synset into consideration. In this algorithm, confidence is not specifically the statistical meaning, rather it is an estimation for how certain DARCI is about a particular synset.

$$\text{score} = o * \left[(p+n) * min\left(1, \frac{n}{p}\right)\right]^{\left(\frac{o-0.5}{\gamma}\right)} \tag{1}$$

Here o is the output of a neural network for a particular synset, p is the number of positive data instances present in the training database and n the number of negative data instances, and γ is a constant that indicates how much effect the "confidence" measure should have—we found $\gamma = 5$ to be useful. This equation amplifies outputs of synsets with greater support $(p + n)$ and at least as many negative as positive examples (there would be more negative than positive examples in an accurate sample of the real world). It is immediately clear that synsets having no negative examples will have a score of zero, thus preventing overly positive data from tainting the labeling process.

DARCI then uses this modified score to make her selection of synset labels with the added caveat that no two synsets are chosen that belong to the same satellite group of synsets. Satellite groups are groupings of adjective synsets defined in WordNet to share similar meanings. It is a grouping that is looser than the synset grouping itself, but still somewhat constrained. For example, all colors belong to the satellite group "chromatic". This means that DARCI will never label an image with more than one color. We do this in order to enforce a varied selection of labels.

6 Results

Because labeling images with adjectives is subjective, it is difficult to evaluate DARCI's progress. And since DARCI is not yet producing any artefacts, we can't directly assess how the associations she is currently learning will effect those artefacts. Nevertheless, in this section we present the results of a test that we devised to estimate how DARCI is learning select adjectives, with the caveat that the evaluation is still somewhat subjective. We also demonstrate DARCI's labeling capabilities for a handful of images. Finally, we briefly describe DARCI's ability to select the top images, from our database, that fit a given adjective label.

As of this writing, there were 1284 images in our image database and a total of 5891 positive user provided labels. 3465 of those labels belonged to synsets that passed the requirement of eight minimum labels. There were 150 synsets that passed this requirement, constituting the synsets that we say DARCI knows. Even though the system is designed to update incrementally, we re-ran all of the data from scratch using updated parameters.

6.1 Empirical Results

In order to assess DARCI's ability to associate words with image features, we observed DARCI's neural net outputs for ten select synsets across ten images that were not in our image database. We presented these same images and synsets to online users in the form of a survey. We chose this narrow survey approach for evaluation because the data available for each image in our labeled dataset was scarce. On the survey, users were asked to indicate whether or not each word described each image. They were also given the option to indicate *unsure*. Across the ten images, each synset received 215 total votes. For every synset, the positive count for each image was normalized by the total number of votes that the image received for the given synset. We then calculated the correlation coefficient between DARCI's neural network output and this normalized

Synset	Gloss	Correlation Coefficient	p-value
Scary	provoking fear terror	0.1787	0.6214
Dark	devoid of or deficient in light or brightness; shadowed or black	0.7749	0.0085
Happy	enjoying or showing or marked by joy or pleasure	0.0045	0.9900
Sad	experiencing or showing sorrow or unhappiness	0.3727	0.2888
Lonely	lacking companions or companionship	0.4013	0.2504
Wet	covered or soaked with a liquid such as water	0.3649	0.2998
Violent	characterized by violence or bloodshed	0.2335	0.5162
Sketchy	giving only major points; lacking completeness	0.4417	0.2013
Abstract	not representing or imitating external reality or the objects of nature	0.2711	0.4486
Peaceful	not disturbed by strife or turmoil or war	0.3715	0.2905

Table 1. Empirical results over ten synsets across ten images. The gloss is the WordNet definition. The correlation coefficient is between DARCI's neural net outputs and normalized positive votes from humans. The p-value is for the correlation coefficient.

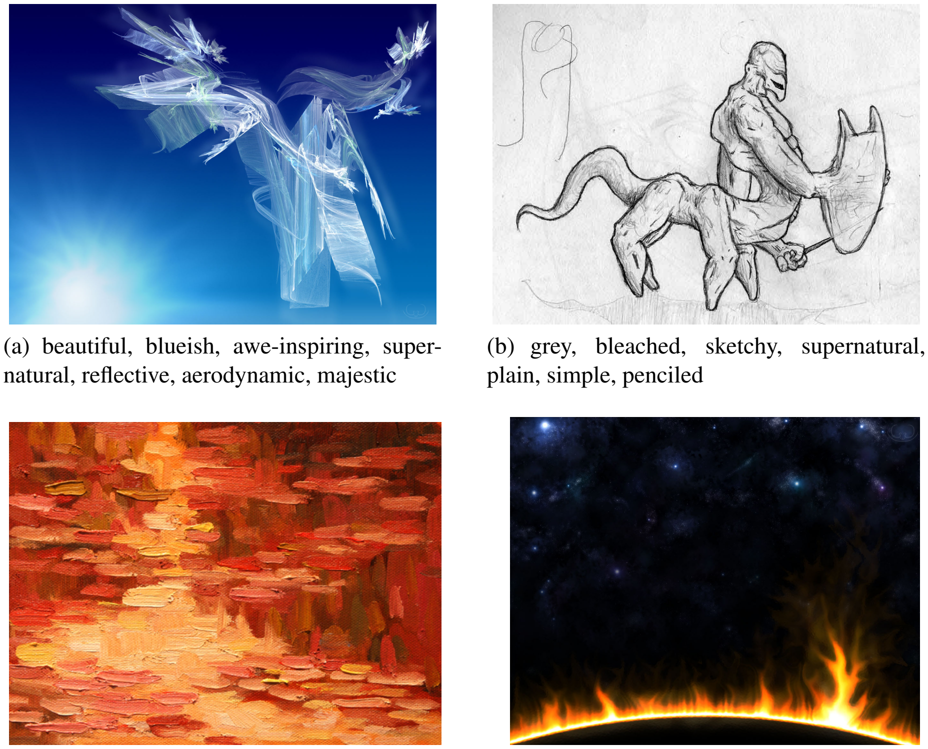
positive count. Table 1 shows the results of this experiment for each synset along with the accompanying p-value.

A high positive correlation and a statistically significant p-value would indicate that DARCI agrees with the majority of those surveyed. The p-values we obtained indicate, unfortunately, that for the most part, these results are not statistically significant. However, all of the synsets have a positive correlation, hinting that the system is heading in the right direction and had we more data, would probably be significant. Of note is the synset "dark", which has the highest correlation coefficient and is statistically significant to $p < 0.01$. "Happy" is both the least statistically significant and shows essentially no correlation between DARCI's output and the opinions of users. From these results, and acknowledging the small amount of training data we have acquired, we can surmise that DARCI is capable of learning to apply some synsets quite effectively, while other synsets may be impossible for DARCI to learn. More data will be necessary to solidify these conjectures.

It is important to note that humans don't always interpret images in the same fashion themselves. For example, the results regarding the synsets for "sketchy", "sad", and "lonely" showed little agreement amongst the human participants. While disagreement amongst humans did not necessarily correlate with DARCI's interpretations, the subjectivity of the problem somewhat absolves DARCI of the necessity for high correlation with common consent among humans. Clearly, other metrics are needed to truly evaluate DARCI.

6.2 Anecdotal Results

We presented DARCI with several images that were not in her database, and observed her descriptive labels of them. Figure 2 shows some of the images and the seven adjectives that DARCI used to describe them. In this figure we see that DARCI did fairly well in describing these four images. Though subjective, a case can be made for describing



(a) beautiful, blueish, awe-inspiring, supernatural, reflective, aerodynamic, majestic

(b) grey, bleached, sketchy, supernatural, plain, simple, penciled

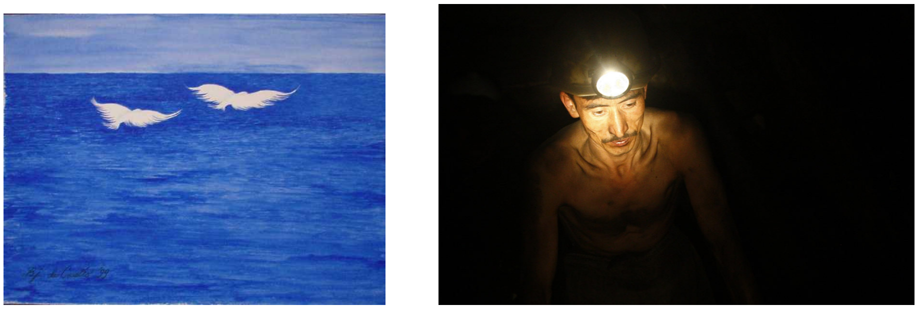
(c) orange, hot, supernatural, painted, blotched, abstract, rough

(d) scary, beautiful, violent, dark, red, fiery, supernatural

Fig. 2. Images that DARCI has interpreted. Words underneath each image are the adjectives DARCI associated with each image. (a), (b), and (d) courtesy of Shaytu Schwandes. (c) courtesy of William Meire.

each image the way DARCI did. One exception would be the adjective "supernatural" which appears in every single image DARCI labels. Until DARCI sees enough negative examples of "supernatural", she will continue learning that all images are "supernatural" because she has mostly seen only positive examples of the word.

DARCI's vocabulary, as of now, is 150 adjective synsets and she has learned some synsets better than others based on two things. First, she has seen more examples of some synsets than of others. Second, some synsets are simply much more difficult to learn. For example, for DARCI to determine whether an image is "dark" or not is much easier than for her to determine whether or not an image is "awesome". "awesome" is much more subjective and takes more aspects of the image into consideration. DARCI had never seen the images shown in Figure 2 before and so, to analogize with the human process, she had to describe the images based on her own experience. One could argue that DARCI was showing imagination because she came up with appropriate adjectives.



(a) peaceful

(b) lonely

Fig. 3. Representative images that DARCI listed in her top ten images described as (a) peaceful and (b) lonely. These images were not explicitly labeled as such when they first appeared in these lists. (a) courtesy of Bj. de Castro. (b) courtesy of Ahmad Masood.

We designed DARCI so that she could find and display the top ten images she thinks are described by a particular adjective synset as well as the top ten images she thinks do not represent that particular synset. This gives us a good idea of how well DARCI has learned a particular synset. It is interesting to note that images that have not been explicitly labeled with a particular synset often show up in DARCI's lists. In Figure 3 we see two examples of this with the adjectives "peaceful" and "lonely". DARCI displayed these two images as respectively "peaceful" and "lonely" even though they had never been explicitly labeled as such. Many would agree that these two images are in fact describable as DARCI categorized them. Again, one could argue that DARCI was showing imagination because she displayed these images on her own. To observe DARCI's image interpreting capabilities go to her website [1].

7 Discussion and Future Work

In this paper we have outlined and demonstrated the first critical component of DARCI. This component is responsible for forming associations between image features and descriptive words, and represents an aspect of artistic appreciation that is critical for the next steps in DARCI's development. The next component will be responsible for rendering images in an original and aesthetically pleasing way that reflects a series of accompanying adjectives. For example, we may present DARCI with a photograph of a lion and the words: majestic and scary. DARCI would then create an artistic rendering of the lion in a way that conveys majestic and scary. If DARCI is able to learn how to render images according to any combination of descriptive words, then the possibility for original and meaningful art becomes apparent. The argument for creativity is strengthened as well. For example, what if one were to commission DARCI to render the photograph of a forest scene in a way that is photographic, abstract, angry, and calm? Who could say what the final image would look like? The commissioner may be attributed with creativity for coming up with such a contradictory set of words, but the greater act of creativity would arguably lie in the hands of DARCI.

The rendering component of DARCI will use a genetic algorithm to discover how to render images in a way that reflects accompanying adjectives. The fitness function for this algorithm will be largely a measure of how closely the phenotype, a rendered image, matches the adjective in question. This measure will be the very output of the adjective's associated neural net described in the body of this paper—it is a measure of her appreciation for her own work. Since DARCI is persistent, this means that the fitness function will be changing as her associative abilities improve. In fact, we intend to introduce some of her own images into the database, thus convolving the associative and productive processes. For this reason, we want DARCI to strengthen her associations *while* she produces and evaluates her own images.

Once the rendering component of DARCI is complete, we will continue to develop her ability to be creative. We intend to allow DARCI to select the adjectives that drive image creation by some process that takes associative knowledge into consideration. We may form associations between adjectives and nouns/verbs. This would provide a framework for DARCI to choose the subjects to render based on image captions. Finally, we hope to eventually allow DARCI to create images from scratch, prior to rendering, using a cognitive model that would rely heavily on the associative component.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. IIS-0856089.

References

1. DARCI (Digital ARtist Communicating Intention). http://axon.cs.byu.edu/DARCI/.
2. Distributed content-based visual information retrieval system on peer-to-pear(p2p) network. http://appsrv.cse.cuhk.edu.hk/~miplab/discovir/.
3. S. Colton. Creativity versus the perception of creativity in computational systems. *Creative Intelligent Systems: Papers from the AAAI Spring Symposium*, pages 14–20, 2008.
4. R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. *Lecture Notes in Computer Science*, 3953:288–301, 2006.
5. C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. The MIT Press, 1998.
6. T. Gevers and A. Smeulders. Combining color and shape invariant features for image retrieval. *IEEE Transactions on Image Processing*, 9:102–119, 2000.
7. C. Li and T. Chen. Aesthetic visual quality assessment of paintings. *IEEE Journal of Selected Topics in Signal Processing*, 3:236–252, 2009.
8. X. Shen, M. Boutell, J. Luo, and C. Brown. Multi-label machine learning and its application to semantic scene cassification, 2004.
9. W.-N. Wang and Q. He. A survey on emotional semantic image retrieval. *Proceedings of the International Conference on Image Processing*, 2008.
10. W.-N. Wang, Y.-L. Yu, and S.-M. Jiang. Image retrieval by emotional semantics: A study of emotional space and feature extraction. *IEEE International Conference on Systems, Man, and Cybernetics*, 4:3534–3539, 2006.
11. J. Zujovic, L. Gandy, and S. Friedman. Identifying painting genre using neural networks. *miscellaneous*, 2007.