Communicating Affective Meaning from Software to Wetware Through the Medium of Digital Art

R David Norton
Brigham Young University - Provo

Follow this and additional works at: https://scholarsarchive.byu.edu/etd

Part of the Computer Sciences Commons

BYU ScholarsArchive Citation
https://scholarsarchive.byu.edu/etd/5546

This Dissertation is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in All Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.
Communicating Affective Meaning from Software to Wetware
Through the Medium of Digital Art

R. David Norton

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Dan Ventura, Chair
  Michael Jones
  Michael Goodrich
  Bryan Morse
  Dennis Ng

Department of Computer Science
Brigham Young University
August 2014

Copyright © 2014 R. David Norton
All Rights Reserved
ABSTRACT

Communicating Affective Meaning from Software to Wetware Through the Medium of Digital Art

R. David Norton
Department of Computer Science, BYU
Doctor of Philosophy

Computational creativity is a new and developing field of artificial intelligence concerned with computational systems that either autonomously produce original and functional products, or that augment the ability of humans to do so. As the role of computers in our daily lives is continuing to expand, the need for such systems is becoming increasingly important. We introduce and document the development of a new “creative” system, called DARCI (Digital ARtist Communicating Intention), that is designed to autonomously create novel artistic images that convey linguistic concepts to the viewer. Within the scope of this work, the system becomes capable of creating non-photorealistic renderings of existing image compositions so that they convey the semantics of given adjectives. Ultimately, we show that DARCI is capable of producing surprising artifacts that are competitive, in some ways, with those produced by human artists.

As with the development of any “creative” system, we are faced with the challenges of incorporating the philosophies of creativity into the design of the system, assessing the system’s creativity, overcoming technical shortcomings of extant modern algorithms, and justifying the system within its creative domain (in this case, visual art). In meeting these challenges with DARCI, we demonstrate three broad contributions of the system: 1) the contribution to the field of computational creativity in the form of an original system, new approaches to achieving autonomy in creative systems, and new practical assessment methods; 2) the contribution to the field of computer vision in the form of new image features for affective image annotation and a new dataset; and 3) the contribution to the domain of visual art in the form of mutually beneficial collaborations and participation in several art galleries and exhibits.

Keywords: Computational Creativity, Genetic Algorithms, Artificial Neural Networks, Affective Image Annotation, Image Features, Visual Art
ACKNOWLEDGMENTS

I would like to first thank my adviser, Dr. Dan Ventura, for his continual support, guidance, and willingness to let me forge my own path through the unusual discipline of computational creativity. I would also like to thank the rest of my committee for the time and consideration they have given to my research. I would especially like to thank my collaborator, Derrall Heath, for his insights and assistance throughout the development of DARCI. I wish him luck as he continues making DARCI the best it can be. Finally, I would like to thank my dear wife, Stephanie, who has been my greatest supporter during the final and most challenging portions of my research.
Table of Contents

List of Figures ix
List of Tables xiii

1 Introduction 1
   1.1 Challenges of Computational Creativity .......................... 1
   1.2 DARCI ....................................................... 3
   1.3 Background ............................................... 5
   1.4 Scope ...................................................... 8
   1.5 Contributions .............................................. 8
   1.6 Summary of Dissertation ..................................... 9
      1.6.1 Summary of Contribution to Computational Creativity ...... 9
      1.6.2 Summary of Contribution to Computer Vision ............... 11
      1.6.3 Summary of Contribution to the Visual Arts ............... 12

2 Establishing Appreciation in a Creative System 13
   2.1 Introduction ............................................... 14
   2.2 Image Feature Extraction ................................... 15
   2.3 Visuo-Linguistic Association ................................ 16
      2.3.1 WordNet ............................................... 17
      2.3.2 Learning Method ....................................... 17
   2.4 Obtaining Data Instances ................................... 18
      2.4.1 Amplifying Data ....................................... 18
2.4 Interpreting Images ................................................................. 20

2.5 Results .................................................................................... 21

2.6 Empirical Results ............................................................... 22

2.7 Discussion and Future Work .................................................. 26

3  Autonomously Creating Quality Images .................................. 28

3.1 Introduction ........................................................................... 29

3.2 Image Appreciation ............................................................ 30

3.3 Image Creation ................................................................. 31

3.4 Methods and Results ......................................................... 36

3.5 Discussion ........................................................................... 39

3.6 Conclusions .......................................................................... 42

3.7 Acknowledgments ............................................................... 44

4  Finding Creativity in an Artificial Artist ................................. 45

4.1 Introduction ........................................................................... 46

4.2 Background .......................................................................... 47

4.3 Methods ................................................................................. 50

4.3.1 Appreciation ................................................................. 50

4.3.2 Skill and Imagination ...................................................... 53

4.4 Evaluation of Creativity ....................................................... 58

4.4.1 Evaluation of Appreciation ............................................. 58

4.4.2 Evaluation of Skill and Imagination ............................... 60

4.4.3 Effectiveness of Survey ................................................. 63

4.5 Conclusions .......................................................................... 67

5  Autonomously Managing Competing Objectives to Improve the Creation and Curation of Artifacts 69
5.1 Introduction .............................................. 70
5.2 Image Rendering ........................................... 71
  5.2.1 Adjective Matching .................................... 72
  5.2.2 Similarity ............................................. 77
5.3 Experimental Design ....................................... 80
5.4 Results .................................................. 81
  5.4.1 Most Fit Selection ..................................... 85
  5.4.2 Selection After Last Shift ............................... 91
  5.4.3 Filter Sequence Length ................................. 94
5.5 Conclusions .............................................. 95

6 Accounting for Bias in the Evaluation of Creative Computational Systems: An Assessment of DARCI 97

6.1 Introduction .............................................. 98
6.2 Background ............................................... 99
6.3 DARCI and Artifact Creation ............................... 100
  6.3.1 Adjective ANN ......................................... 102
  6.3.2 Genetic Algorithm .................................... 102
  6.3.3 Novelty ANN .......................................... 103
  6.3.4 Candidate Artifact Curation ........................... 104
6.4 Commissions ............................................... 108
  6.4.1 Online Survey ......................................... 110
6.5 Results .................................................. 116
  6.5.1 Bias .................................................. 117
  6.5.2 Evaluation of Creativity ............................... 119
6.6 Conclusions .............................................. 122
7 Improving Affective Image Annotation with Features that Summarize Local Interest Points

7.1 Introduction .............................................. 126
7.2 Background .............................................. 127
7.3 Datasets ................................................... 129
  7.3.1 IAPS .................................................. 129
  7.3.2 Art Photo ........................................... 130
  7.3.3 Abstract ............................................ 131
  7.3.4 DARCI ............................................... 131
7.4 Image Features ............................................ 134
  7.4.1 Color Features ...................................... 134
  7.4.2 Global Texture and Shape Features ............... 142
  7.4.3 Summary of Local Interest Point Features ....... 144
  7.4.4 Feature Selection .................................. 149
7.5 Evaluation ................................................ 152
7.6 Results ................................................... 155
  7.6.1 IAPS Dataset Comparisons ......................... 156
  7.6.2 Art Photo Dataset Comparisons ................... 158
  7.6.3 Abstract Dataset Comparisons .................... 159
  7.6.4 Combined Dataset Comparisons ................... 160
  7.6.5 DARCI Dataset ..................................... 161
  7.6.6 Revised Datasets .................................. 161
7.7 Conclusions ............................................... 164

8 An Artistic Dialogue with the Artificial ..................... 168

8.1 Introduction .............................................. 169
8.2 DARCI ................................................... 171
8.3 Artist Interpretations .................................... 173
## List of Figures

2.1 Screenshot of the website used to train DARCI          .......................................          19
2.2 Images that DARCI has interpreted          .........................................................          24
2.3 Representative images that DARCI has described as peaceful and lonely    ..................          25

3.1 Sample genotype and its effect on an image          ..................................................          32
3.2 The three source images used to evaluate the quality of DARCI’s artifacts  .........................          37
3.3 The average ranking of synsets for the four artifact sources          ......................          39
3.4 Image A rendered “fiery”          .................................................................          41
3.5 Image A rendered “happy”          .................................................................          41
3.6 Image B rendered “happy”          .................................................................          42
3.7 Image B rendered “fiery”, “happy”, and “lonely”          ...........................................          42
3.8 The top “fiery” renderings discovered by DARCI for image A          ..................................          43
3.9 The randomly generated renderings of image A for the synset “fiery”          .................          44

4.1 Sample genotype and its effect on an image          ..................................................          55
4.2 Images that DARCI has interpreted          .........................................................          60
4.3 Images that DARCI created from for ten different adjectives          .................................          63
4.4 Combined survey scores for each adjective DARCI rendered images for          .............................................          65

5.1 The three source images used in all experiments          ...........................................          82
5.2 Sample ‘sad’ images from training data          ......................................................          83
5.3 Sample ‘happy’ images from training data          ......................................................          84
5.4 Sample ‘fiery’ images from training data          ......................................................          85
7.1 One interface for labeling images for the DARCI dataset ................. 132
7.2 One interface for obtaining negative datapoints for DARCI dataset .... 133
7.3 Illustration of aspects of certain image features ......................... 138
7.4 A membership function used to create a color histogram ............... 142
7.5 The flow of forward feature selection .................................. 150
7.6 The flow of backward feature selection .................................. 151
7.7 Separation scores using forward selection on SLIP features ............ 152
7.8 The ranked list of candidate SLIP features for forward selection ...... 153
7.9 Separation scores using forward selection on all features ............... 154
7.10 The ranked list of candidate features for forward selection .......... 155
7.11 Separation scores using backward selection on all features .......... 156
7.12 The ranked list of candidate features for backward selection ......... 157
7.13 The list of features that are selected by both forward and backward selection 158
7.14 Comparison of feature selection approaches .......................... 158
7.15 Comparison of approaches on the IAPS dataset ....................... 159
7.16 Comparison of Norton and Yanulevskaya approaches on the IAPS dataset . 160
7.17 Comparison of approaches on the Art Photo dataset ................... 161
7.18 Comparison of approaches on the Abstract dataset ................... 162
7.19 Comparison of approaches on the Combined dataset .................. 163
7.20 Comparison of results on original and revised Art Photo datasets .... 164
7.21 Comparison of results on original and revised Abstract datasets ....... 165
7.22 Comparison of results on original and revised datasets ................. 166
7.23 Comparison of results on original and revised DARCI datasets ....... 167

8.1 Example of DARCI in action ........................................... 170
8.2 Overview of DARCI’s artifact creation process ....................... 173
8.3 The instructions for System 1 ........................................... 177
8.4 The results of the first round of System 1 ............................ 178
8.5 The results of the second round of System 1 .......................... 179
8.6 The results of the third round of System 1 .............................. 179
8.7 Fourth round and final image results for System 1 .................. 180
8.8 The instructions for System 2 .............................................. 181
8.9 The results of System 2 ..................................................... 182
8.10 Three example images created during the Bromerly activity ...... 183
8.11 The cluster of flamping (lonely) images ............................... 183
8.12 The cluster of orfly (weird) images ................................... 184
8.13 The cluster of bromerly (blocky) images .............................. 184
8.14 Images produced by a human and DARCI for the adjectives bromerly (blocky),
      blemistic (majestic), and diambic (avian) ......................... 186
8.15 Images produced by a human and DARCI for the adjectives flamping (lonely)
      and orfly (weird) .................................................... 186
8.16 Images produced by a human and DARCI for the adjectives orfly (weird),
      blemistic (majestic), and diambic (avian) ......................... 187
8.17 The template used to evaluate the objective group .................. 191
8.18 Examples of drawings created by the group evaluated subjectively . 192
8.19 Examples of drawings created by the group evaluated objectively . 193
8.20 The ten images that received jury awards at Fitness Function ...... 194
8.21 Two sample submissions from a witty artist ......................... 195

9.1 Fitness Function art exhibit .............................................. 199
9.2 Submissions to Utah County Art Gallery .............................. 200
9.3 Submissions to Evolutionary Art, Design, and Creativity Competition . 201
9.4 A portrait produced by The Painting Fool in collaboration with DARCI 203
9.5 The Painting Fool, in collaboration with DARCI, refusing to paint .... 204
# List of Tables

2.1 Image features extracted from images for association with adjectives . . . . . 16
2.2 Correlation between neural net outputs and human opinion . . . . . . . . . 23
3.1 Parameters used for the evolutionary mechanism . . . . . . . . . . . . . . 35
3.2 The average ranking of artifacts . . . . . . . . . . . . . . . . . . . . . . . . 39
3.3 Statistical analysis of result differences . . . . . . . . . . . . . . . . . . . 40
4.1 Image features extracted from images for association with adjectives . . . . 51
4.2 Parameters used for the evolutionary mechanism . . . . . . . . . . . . . . 57
4.3 Results of appreciation network compared with a random untrained ANN . . 59
4.4 The average score of each adjective for each survey item . . . . . . . . . . . 64
4.5 Alpha values measuring consistency of survey questions . . . . . . . . . . 65
4.6 Alpha values measuring consistency of volunteer responses to images . . . . 66
5.1 Image features used to train neural networks . . . . . . . . . . . . . . . . . . 74
6.1 Correlations between volunteer bias and scores for informed experiments . . 118
6.2 Correlations between volunteer bias and scores for blind experiments . . . . 119
6.3 The four Likert scores for the source image . . . . . . . . . . . . . . . . . . 120
6.4 Top six ranked images for each Likert item for the informed experiment . . 121
6.5 Top six ranked images for each Likert item for the blind experiment . . . . . 122
7.1 Statistics about the DARCI dataset . . . . . . . . . . . . . . . . . . . . . . . 134
7.2 Summary of features from prior work . . . . . . . . . . . . . . . . . . . . . 135
Chapter 1

Introduction

As computers become more powerful and ubiquitous in society, the role of artificial intelligence becomes increasingly important. Artificial intelligence, the computational automation of any task requiring intelligence, has found its way into everyday applications from information retrieval to driving cars. As our expectations of electronic applications grow, so too does the need for a more thorough simulation of intelligence. Creativity is an attribute often associated with intelligence [38]. While there is no universal agreement on what creativity is [48], most definitions include the concepts of novelty and usefulness [65]. Computational creativity is a new and developing field of artificial intelligence concerned with computational systems that either autonomously produce novel and useful products, or that augment the ability of humans to do so. Such products have included pieces of art [13, 61], music [19, 46, 63, 93], video games [18], culinary recipes [64, 88], jokes [4], poems and stories [34, 75, 90], metaphors [89], mathematical theories [11], and so forth.

1.1 Challenges of Computational Creativity

All research in computational creativity, involving the development of new systems, faces three fundamental types of challenges. The first type of challenge is that inherent to the domain of computational creativity itself. Creativity is arguably what Gallie described to be an essentially contested concept [31]. Jordanous paraphrased Gallie’s description saying that an essentially contested concept is “a subjective concept whose meaning seems to be commonly understood, with a variety of interpretations available to be attached to that
concept, but where a fixed ‘proper general use’ is elusive” [48]. Essentially, there is no one “correct” philosophy of what it means to be creative. Thus, researchers in computational creativity must pick the philosophies they subscribe to and design and validate their systems accordingly.

Central to the discussion of creativity in the field of computational creativity are two fundamental viewpoints regarding the design of creative systems. These viewpoints are designated as weak computational creativity, emphasizing the artifacts produced by creative systems, and strong computational creativity, emphasizing the process by which the artifacts are produced [1]. These two viewpoints allude to two specific problems facing computational creativity as a whole: system autonomy and system evaluation. Finding the right amount of system autonomy for any given research goal is the first of these problems. A high degree of autonomy is efficient (as human intervention is usually costly) and arguably necessary to attribute creativity to the system itself, but it often leads to a degradation in artifact quality [17]. The second problem is assessing the creativity of a given system so that research progress can be measured and different systems can be compared to one another. Research on system evaluation has yielded many valuable assessment approaches [49]. Unfortunately, none of these approaches have yet to become discipline standards.

The second type of challenge facing computational creativity research is that inherent to producing any state-of-the-art computer system. Producing systems that perform intelligent and creative tasks often requires the combination of algorithms spanning many disciplines of computer science such as machine learning, natural language processing, information retrieval, signal processing, and computer vision. Often, extant algorithms are applied in novel ways that require extensive modification or expansion. Additionally, entirely new algorithms with broad application are commonly developed to achieve the desired system behavior.

The third type of challenge facing computational creativity research is that inherent to the task of producing artifacts in a creative domain. Ultimately, the systems created for computational creativity research must be justified within their given creative domain, be it...
music, visual art, mathematics, etc. This justification is independent of any merit the systems may have achieved within the discipline of computational creativity, or even computer science at large. In Csíkszentmihályi’s systems model of creativity, he argues that for an idea to be culturally creative it “... must be couched in terms that are understandable to others, it must pass muster with experts in the field, and finally it must be included in the cultural domain to which it belongs” [21]. Ideally, a creative system would pass these social requirements.

1.2 DARCI

In this dissertation we present a new computational system we have developed to explore computational creativity in the domain of visual art. This system, called DARCI (Digital ARtist Communicating Intention), is designed to autonomously produce original and aesthetically stimulating digital images that communicate meaning to the viewer. In developing DARCI, we have subscribed to several philosophies of creativity spanning a variety of viewpoints.

Creativity is traditionally considered to involve the act of producing something, an artifact, that is both novel and useful [65]. Aligned with this traditional notion, Boden has proposed the requirements of novelty and quality in attributing creativity to a computational agent [6]. Novelty is the notion that an artifact, or process, is original within some scope, and quality is the notion that an artifact is of some value to others, be it practical, aesthetic, or otherwise. Two variations on this theme that we have consented to are Colton’s creative tripod [12], and Richie’s essential properties of creativity [77].

Colton’s creative tripod focuses on the process of creativity (strong computational creativity), as opposed to the product, with a special emphasis on the perception of creative behavior [12]. Here quality is replaced by skill, the capacity of a system to produce quality artifacts; and novelty is replaced by imagination, the capacity of the system to produce original and meaningful (non-random) artifacts. Added to Boden’s staple attributes is appreciation, the ability of the system to recognize the quality and novelty of its own artifacts (reinforcing the emphasis on process over product). Ritchie focuses on the artifacts produced by the
system (weak computational creativity) and adds typicality to Boden’s two criteria [77]. As Ritchie defines it, typicality is the extent to which an artifact is a member of its intended domain. In the domain of visual arts for example, typicality may be the extent to which an artifact is a visual depiction of a scene or idea.

Csikszentmihalyi’s systems model of creativity, mentioned previously, describes creativity as a social phenomena consisting of the domain, the field, and the person [21]. The person is the entity performing the creative act, the field is the group of experts for a particular creative domain, and the domain is the set of rules by which the field evaluates a body of work. The systems model of creativity is focused on identifying the social interaction of creative processes and the social impact of creative artifacts rather than the explicit definition of creativity. Csikszentmihalyi’s model allows for the definition of creativity to change over time as persons, fields, and domains evolve.

Fundamental to the conception of DARCI is the notion that art must elicit an aesthetic experience in the viewer and that “the aesthetic experience occurs when information coming from the artwork interacts with information already stored in the viewer’s mind” [22]. In order for the system to produce images that contain information that can interact with the human mind, DARCI must be able to associate linguistic concepts with images and then reproduce those concepts. We have designed DARCI to learn these associations from interactions with human subjects in order to conform with Csikszentmihalyi’s systems model creativity. However, in order to increase autonomy, DARCI is designed to produce artifacts independently from human interaction. This autonomous creation process is designed to adhere to Colton’s creative tripod and to produce artifacts that satisfy Ritchie’s essential properties of creativity. Furthermore, we argue that the concepts that DARCI’s artifacts are intended to communicate represent the system’s intention. Validating the importance we have placed on intentionality and social interaction, Jordanous has recently included these two attributes in her fourteen components of creativity, a formalism of creativity now recognized in the field of computational creativity [47].
1.3 Background

There have been many computational systems designed to autonomously produce visual art. One of the most prominent systems recognized in the domain of computational creativity is Harold Cohen’s AARON [61]. While AARON has succeeded in producing pieces of art recognized and praised by the field of visual art, AARON’s creation mechanism has been almost entirely concealed from the public. Where Cohen’s influence ends and where AARON’s begins is unknown.

While the mechanisms governing AARON are concealed, other image generating systems have been openly described in the computational creativity community. Many of these systems, including DARCI, employ evolutionary mechanisms due to the innate ability of evolution to yield unpredictable solutions to problems [33].

Traditionally, this evolutionary art has been an interactive affair involving the creation of a virtual space of myriad possible images for human observers to explore. This space of possible images is defined by a digital encoding, or genotype, and the rules to transform this genotype into an actual image, commonly called the phenotype. Evolutionary mechanisms begin with a population of random genotypes which are evaluated based on the qualities of their respective phenotypes. This evaluation is effected by a fitness function that determines which genotypes get to pass on their traits to future generations of the evolutionary mechanism. Since art is inherently subjective and difficult to parameterize, this evaluation is often left to human judgment. Hence a human’s aesthetic sense acts as the fitness function for most evolutionary computation used for producing art. Sims was one of the first researchers to propose how such an evolutionary mechanism would work in the domain of visual art [83].

Involving human judgment in the evolutionary process is time consuming; but more importantly, it is an unsatisfactory component to many prospective creative systems since it removes some of the autonomy desired in such systems. Several evolutionary systems have been developed to automatically assign a fitness function to phenotypes. Most of these systems extract quantifiable features from the images to evaluate their aesthetics.
Examples of features used in existing evolutionary systems include: how closely the image color distribution matches the color distribution of highly rated Flickr images [73], how complex the image is as measured by image compression [56], a geometric assessment of regions within an image [36], and so forth.

More sophisticated approaches use a dynamic fitness function that changes based on conditions in the evolutionary environment. For example, DiPaola and Gabora have designed a system that examines two sets of features when assigning fitness to images: one is the degree to which an image is similar to a target image and the other is the degree to which an image follows quantifiable rules of aesthetics. The weight that the algorithm puts on each set of features changes whenever evolution becomes stagnant [27]. Another example of a dynamic fitness function is a co-evolutionary model implemented by Greenfield. In this algorithm, the system co-evolves a population of images and a population of image filters together. The fitness functions of each population differ, but are dependent upon the interaction between the two populations. Thus, the fitness functions fluctuate as if in an ecological setting [37].

Another approach to automating fitness evaluation that has been explored is to model natural systems that can be used to evaluate images. For example, one could model a human’s appreciation of images and use that model as the fitness function rather than an actual human. This particular case was first explored by Baluja et al. using an artificial neural network to model the aesthetic preferences of specific users [2]. To train the neural network, Beluja used data collected from a traditional human-as-the-judge evolutionary mechanism. The entire pixel space of each image was the input for the neural network. Beluja had minimal success, noting the difficulty in using such an enormous input space for such a sparse selection of data points.

Recently, Machado et al. augmented their NEvAr system by incorporating a dynamic artificial neural network model to evaluate the novelty of the system’s creations [57]. They trained their new model to distinguish between images that NEvAr created and paintings by renowned artists. As suggested by Beluja, for the neural net inputs they used carefully
selected image features rather than the entire pixel space. Machado then used the model to act as the fitness function for NEvAr’s evolutionary mechanism. Images that the system recognized as its own creations were rejected, thus the system had to evolve to change its own style. After terminating the evolutionary mechanism, the rejected creations were added to the neural network model as NEvAr’s creations. This process was repeated twelve times, forcing NEvAr to dramatically change its style.

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

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

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

To conform with Csíkszentmihályi’s systems model of creativity, DARCI is designed to learn and create within a social context primarily through the its interactions with people via the internet. Ultimately, DARCI’s representation of concepts is a reflection of those who have participated in training the system. As previously mentioned, traditional evolutionary mechanisms require social interaction as well; however, these traditional algorithms involve the explicit creation of artifacts as opposed to the creation of a language—in DARCI’s case, a language for communicating concepts with images.
1.4 Scope

DARCI is ambitious in scope, and while the design for DARCI calls for the system to be able to communicate any linguistic concept in a fully composed original image, for this dissertation, DARCI’s capabilities are limited to communicating adjectives in non-photorealistic renderings of previously composed images. To do so, DARCI employs a series of neural networks and human labeled training data to learn how to associate image features with any adjective. These neural networks in turn act as the fitness function for a type of evolutionary mechanism called a genetic algorithm. This genetic algorithm discovers sequences and parameter settings of non-photorealistic image filters (like those used in Photoshop, Gimp, or other photo-editing software) that will modify a specified pre-existing image so that it will reflect a specified list of adjectives. Future work will extend DARCI’s capabilities to the full design goals.

1.5 Contributions

In developing DARCI, we have met the three types of challenges facing research in computational creativity outlined earlier. As such, DARCI’s contributions are three-fold.

First, DARCI’s particular emphasis on intention and social interaction within the domain of visual art are unique to the field of computational creativity [70]. The development of DARCI has spanned several years over which features have been gradually added, system autonomy increased, and practical evaluation methods developed and improved. In Chapters 2 through 6 we document this progression and illustrate how specific developments are generalizable to the broader field of computational creativity.

Second, the development of DARCI has necessitated improvements to state-of-the-art affective image annotation techniques within the discipline of computer vision. These improvements, along with a new dataset, are described in Chapter 7.
Third, DARCI has influenced individuals in the domain of visual art through a mutually beneficial collaboration described in Chapter 8 and participation in several art galleries and exhibits summarized in Chapter 9.

### 1.6 Summary of Dissertation

This dissertation is divided into three parts, each of which covers our contribution regarding one of the three types of challenges facing computational creativity. A summary of the chapters pertaining to each contribution is detailed below.

#### 1.6.1 Summary of Contribution to Computational Creativity

In Chapter 2 we introduce the first version of DARCI’s image-adjective association component—that is, a program for labeling images with adjectives\(^1\). By conducting a simple online survey, we demonstrate how DARCI will be able to evaluate its own images in later chapters.

In Chapter 3 we demonstrate a functional creative system\(^2\). We describe in detail how DARCI uses the image-adjective association component of Chapter 2 to govern a genetic algorithm that discovers non-photorealistic renderings of pre-existing images that convey given adjectives. We conduct a detailed online survey asking participants to rank artifacts by how well they represent desired adjectives. These artifacts include those produced by DARCI, those produced by human artists, those produced as a collaboration between DARCI and humans, and those selected from random renderings. We show that DARCI is able to produce renderings that convey adjectives to an audience comparably with renderings produced by human artists. Furthermore, we show that, for the given task, a human/DARCI collaboration can achieve results superior to either acting in isolation.

In Chapter 4 we analyze the system described in Chapter 3 in terms of Colton’s creative tripod\(^3\) [12]. We also assess the “creativity” of DARCI’s artifacts for the first time.

---

\(^1\)Published in the proceedings of the 1st International Conference on Computational Creativity
\(^2\)Published in the proceedings of the 2nd International Conference on Computational Creativity
\(^3\)Published in the Journal of Creative Behavior
using an original (and generalizable) online survey composed of psychologically inspired Likert items. We are able to validate the survey by showing that the items are statistically consistent and correlate with participants’ perceptions of creativity. Though we don’t have a baseline in this study for comparison, we show that for some images, viewers elicit overall positive responses.

In Chapter 5 we compare and contrast variants of the fitness function used by DARCI’s genetic algorithm\(^4\). These variants include a new metric that calculates the similarity of an artifact to the original image (pre-rendering). We perform a qualitative evaluation to determine the most effective fitness function for DARCI. We also demonstrate a fully autonomous DARCI by publishing almost every artifact (the few exceptions are left out for space and noted) produced by DARCI for these experiments.

In Chapter 6 we further enhance the autonomy of DARCI with a meta-process that essentially creates a pool of candidate artifacts with a progressive genetic algorithm\(^5\). This meta-process, while original in our use, is inspired by the work of Machado et al. \([57]\). In addition to this new meta-process, we incorporate a new and improved image-adjective association component described in Chapter 7. The resulting system represents the final state of DARCI for this dissertation. We assess the “creativity” of DARCI using an online survey comparable to the one in Chapter 4. In this new survey we consider the process of creation in addition to the artifacts produced. We also consider evaluator bias and compare DARCI’s artifacts to those produced by human artists. We show that DARCI produces artifacts that score statistically the same as human artifacts on Likert items pertaining to how much participants like the renderings, how closely they match desired adjectives, and how well the renderings make use of a source photograph. We show that DARCI produces artifacts that score statistically higher than human artifacts on Likert items pertaining to

\(^4\)Published in the proceedings of the 5\(^{th}\) International Conference on Computational Creativity

\(^5\)Reserved for submission to the proceedings of the 6\(^{th}\) International Conference on Computational Creativity
how surprising the renderings are and how difficult the participants perceive the process of creation to be. DARCI does not score statistically lower than humans on any items.

1.6.2 Summary of Contribution to Computer Vision

In order to form image-adjective associations, DARCI needs to be able to analyze digital images and classify them appropriately. In computer vision, this task most closely belongs within the purview of affective image annotation. While there is research being done in this area, most is limited to an oversimplified set of labels with inadequate results. For example, Datta et al. and Li et al. assess the aesthetic quality of respectively photographs and paintings, both limiting their annotation of images to a single affective descriptor, aesthetic, while ignoring the multitude of other potential affective descriptors for images. Wang et al. learn to identify 12 dichotomous pairs of emotional words using specially designed global features [92], Zujovic et al. classify images of paintings into one of six different genres [97], and both Yanulevskaya and Machajdik et al. classify images with eight psychologically derived emotions [58, 94]. Again, all of these approaches limit the classification of images to a very small set of affective descriptors. Furthermore, with the exception of Machajdik et al., none of these researchers make their experimental datasets available, inhibiting other researchers from assessing the value of their own approaches.

In Chapter 7 we go into greater detail regarding these inadequacies and introduce novel image features that improve upon the state-of-the-art6. In addition, we introduce and justify a new dataset to the computer vision community—a dataset consisting of image-adjective associations obtained from human interactions with DARCI. Finally, we demonstrate the superiority of our affective image annotation approach with an extensive comparison to several state-of-the-art techniques using three established datasets.

---

6Submitted to the IEEE Transactions on Affective Computing
1.6.3 Summary of Contribution to the Visual Arts

The bulk of this dissertation is dedicated to scientific discovery and advancement; however, DARCI has also had an impact on individuals in the arts. In 2010, we collaborated with Brigham Young University’s Visual Arts program to hold an advanced Visual Arts course built around DARCI. The course was designed to help Visual Arts students better understand their own creative process and to help us better understand how human artists perform their craft. The collaboration culminated in an art exhibit, called *Fitness Function*, curated by DARCI in the Harris Fine Arts Center at BYU. In Chapter 8 we summarize the collaboration as a case study⁷.

In 2011, we submitted two images created with DARCI to the Utah County Art Gallery: Fall Photography and Digital Art Show, one of which won 2nd place in the Digital Art category. Later that year, we repeated a condensed version of *Fitness Function* in The High Museum of Art in Atlanta, Georgia, as part of the 2011 Conference on Creativity and Cognition. This event was covered by Studio 360 on Public Radio International⁸. DARCI was part of two additional art exhibits in 2013. The first, was the GECCO 2013 Evolutionary Art, Design, and Creativity Competition held in Amsterdam. The second was in collaboration with Colton’s *The Painting Fool* at a public gallery held in Paris. These cultural showcases of DARCI are summarized in Chapter 9.

---

⁷Published in the proceedings of the 8th ACM Conference on Creativity and Cognition—recipient of Best Presentation Award

⁸http://www.studio360.org/story/175741-darci-computer-great-taste/
Chapter 2

Establishing Appreciation in a Creative System

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.

2.1 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, Wiggins, 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 [12], 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, the system needs to first acquire some basic understanding of art [12]. For example, in order for DARCI to appreciate an image that is gloomy, it 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, the system 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 its production of original images.
2.2 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) [50][35]. 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 [24][53].

Wang deals with image retrieval specific to emotional semantics [92][91]. 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 [96]. 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 shown in Table 2.1.
Table 2.1: Image features extracted from images for association with descriptive words.

<table>
<thead>
<tr>
<th>Color &amp; Light:</th>
<th>Texture:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Average red, green, and blue</td>
<td>1. Co-occurrence matrix (x4)</td>
</tr>
<tr>
<td>2. Average hue, saturation, and intensity</td>
<td>2. Maximum probability</td>
</tr>
<tr>
<td>3. Saturation and intensity contrast</td>
<td></td>
</tr>
<tr>
<td>4. Unique hue count (from 20 quantized hues)</td>
<td>3. First order element difference moment</td>
</tr>
<tr>
<td>5. Hue contrast</td>
<td></td>
</tr>
<tr>
<td>6. Dominant hue</td>
<td>4. Entropy</td>
</tr>
<tr>
<td>7. Dominant hue image percent</td>
<td>5. Uniformity</td>
</tr>
<tr>
<td>Shape:</td>
<td></td>
</tr>
<tr>
<td>1. Geometric moment</td>
<td>1. Edge frequency (25x vector)</td>
</tr>
<tr>
<td>2. Eccentricity</td>
<td></td>
</tr>
<tr>
<td>3. Invariant moment (5x vector)</td>
<td>2. Primitive length</td>
</tr>
<tr>
<td>4. Legendre moment</td>
<td>1. Short primitive emphasis</td>
</tr>
<tr>
<td>5. Zernike moment</td>
<td>2. Long primitive emphasis</td>
</tr>
<tr>
<td>7. Edge direction histogram (30 bins)</td>
<td>4. Primitive uniformity</td>
</tr>
<tr>
<td></td>
<td>5. Primitive percentage</td>
</tr>
</tbody>
</table>

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.

2.3 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 the system 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 delimited lists of adjectives.

2.3.1 WordNet

We use WordNet’s [30] 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.

2.3.2 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 [81].

2.4 Obtaining Data Instances

To collect training data, we have created a public website for training DARCI\(^2\). From this website, users are presented with a random image and asked to provide adjectives that describe the image (Figure 2.1). 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.

2.4.1 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.

\(^2http://darci.cs.byu.edu/\)
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 2.1). 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 the system 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.

2.5 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 the system 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 2.1 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.

\[
\text{score} = o \times \left[ (p + n) \times \min \left(1, \frac{n}{p}\right) \right]^{\frac{o-0.5}{\gamma}} \tag{2.1}
\]

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 \(\gamma\) 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 its 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.

2.6 Results

Because labeling images with adjectives is subjective, it is difficult to evaluate DARCI’s progress. And since DARCI is not yet producing any artifacts, we can’t directly assess how the associations the system is currently learning will effect those artifacts. 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.

2.6.1 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 positive count. Table 2.2 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 $p < 0.01$. 

22
Table 2.2: 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.

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

“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.

2.6.2 Anecdotal Results

We presented DARCI with several images that were not in the system’s database, and observed DARCI’s descriptive labels of them. Figure 2.2 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 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”, it will continue learning that all images are “supernatural” because the system has mostly seen only positive examples of the word.

DARCI’s vocabulary, as of now, is 150 adjective synsets and it has learned some synsets better than others based on two things. First, the system 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 it 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 2.2 before and so, to analogize with the human process, the system had to describe the images based on its own experience. One could argue that DARCI was showing imagination because it came up with appropriate adjectives.

We designed DARCI so that it could find and display the top ten images it thinks are described by a particular adjective synset as well as the top ten images it 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 2.3 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 the system displayed these images on its own. To observe DARCI’s image interpreting capabilities, visit the system’s website\(^2\).

Figure 2.3: 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.
2.7 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 the system’s appreciation for its own work. Since DARCI is persistent, this means that the fitness function will be changing as its associative abilities improve. In fact, we intend to introduce some of its own images into the database, thus convolving the associative and productive processes. For this reason, we want DARCI to strengthen its associations while it produces and evaluates its own images.

Once the rendering component of DARCI is complete, we will continue to develop its 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.
Chapter 3

Autonomously Creating Quality Images\textsuperscript{1}

Abstract

Creativity is an important part of human intelligence, and it is difficult to quantify (or even qualify) creativity in an intelligent system. Recently it has been suggested that quality, novelty, and typicality are essential properties of a creative system. We describe and demonstrate a computational system (called DARCI) that is designed to eventually produce images in a creative manner. In this paper, we focus on quality and show, through experimentation and statistical analysis, that DARCI is beginning to be able to produce images with quality comparable to those produced by humans.

3.1 Introduction

DARCI (Digital Artist Communicating Intention) is a computer system designed to eventually create visual art in order to convey intention and meaning to the viewer. Currently, DARCI can automatically render a given image to match an accompanying list of adjectives. This ability is the foundation of a visual language for DARCI to communicate with an audience—an important element of creative expression in the visual arts. DARCI is part of ongoing research that is exploring the perception of creativity in an artificial system.

Measuring creativity both quantitatively and qualitatively is a difficult challenge. Ritchie describes quality, novelty, and typicality as being essential in ascribing creativity to a system [77]. Ritchie defines quality as the extent to which the artifact is a high quality example of its genre. In this paper, we focus on quality, and show that DARCI is beginning to be able to produce quality artifacts comparable to human artists given the same resources.

DARCI’s design has two main components: the image appreciation component, and the image creation component. The image appreciation component is designed to allow DARCI to learn to evaluate it’s own artwork according to various descriptive words. This ability to assess these qualities in an image guides the image creation component. The image creation component uses evolutionary mechanisms to create artifacts and the appreciation component serves as part of the fitness function.

We briefly describe the main components of DARCI and how they work together to produce artifacts. We then present several images that DARCI has created and describe an experiment in which we compare DARCI’s images with ones made by humans. Finally, we discuss how the results show that DARCI is becoming comparable to humans in producing quality artifacts.
3.2 Image Appreciation

It has been argued that the ability to appreciate and evaluate its own artifacts is necessary for a system to be considered creative [12]. In order for DARCI to appreciate art, it must first acquire some basic understanding of art. For example, in order for DARCI to appreciate an image that is dark and gloomy, DARCI must first understand the concepts dark and gloomy. To do this, DARCI must learn to associate images with artistic descriptions.

Image Features

Before DARCI can form associations between images and descriptive words, appropriate image features for the task must be extracted from the image. Significant research has been done in the area of image feature extraction [24, 35, 50, 53, 91, 92], and we have culled 102 image features from this. These are low-level features that can be coarsely classified as treating one the following image characteristics: color, light, texture, and shape.

Artistic Descriptions

As an initial step, the artistic descriptions that DARCI can learn are limited to lists of adjectives. We use WordNet’s [30] database of adjectives to give us a large, yet finite, set of descriptive labels. 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. 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. Additionally, for each image presented to the user, DARCI lists seven adjectives that it associates with the image. The user is allowed to flag those labels that are not accurate. This creates strictly negative examples of those synsets, which is important for learning.

Another program for creatively generating visual art, NodeBox, is also dependent on semantic networks such as WordNet. The NodeBox project takes the use of semantic
networks even further by using a more elaborate database they created called “Perception” [25]. However, unlike DARCI, NodeBox does not have a strong learning component. In the future, we hope to expand DARCI by using more sophisticated semantic networks, perhaps even “Perception” itself.

Learning Method

In order to make the association between image features and synsets, 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. As we incrementally accumulate more data, new neural networks can be dynamically added to the collection to accommodate the new synsets. Currently, there are 211 appreciation networks. This means that DARCI essentially “knows” 211 synsets.

For more details on our learning method, image features, and use of synsets, the reader is referred to earlier work describing DARCI [66].

3.3 Image Creation

DARCI uses a evolutionary mechanism to render images according to given synsets, and this mechanism operates in two modes. The initial mode, which we call practice mode, operates by exploring the space of image filters that will render any image according to a single specific synset. For this mode, DARCI creates and maintains a separate gene pool for each synset that the system knows. The second mode, called commission mode, operates by exploring the space of image filters that will render a specific image according to a specified list of synsets. There is no restriction on synset combinations; in fact, incoherent combinations can produce unexpected and interesting results as we will demonstrate later. For commission mode, users prescribe the image and list of synsets that they wish DARCI to render—in other words, they “commission” DARCI. For each commission, DARCI creates a unique gene pool that
terminates once the commission is complete. For both modes, the evolutionary mechanism functions as follows.

The genotypes that comprise each gene pool are lists of filters, and their accompanying parameters, for processing an image. Many of these filters are similar to those found in Adobe Photoshop and other image editing software. Others come from a series of 1000 filters Simon Colton discovered using his own evolutionary mechanism [16]. Colton’s 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 Colton’s filters as if each category were a unique filter with a single parameter that specifies the specific filter within the category to use. Figure 3.1 gives an example of a genotype and its effect on a sample image. There are a total of sixty-one traditional filters that we selected for DARCI to use and a total of thirty-one categories of filters from Filter Feast, making ninety-two filters available for each genotype. We selected traditional filters that were easily accessible, diverse, fast, and that didn’t incorporate alpha values (since our feature extraction techniques cannot yet process alpha values).
The fitness function for the evolutionary mechanism can be expressed by the following equation:

\[
\text{Fitness}(g) = \lambda_A A(g) + \lambda_I I(g)
\]  

(3.1)

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

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

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

\[
A(g) = \sum_{w \in C} \alpha_w \text{net}_w(g)
\]

(3.2)

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

The interest metric \( I \) penalizes phenotypes that are either too different from the source image, or are too similar. This metric is useful for producing images that meet our definition of imaginative; however, the interest metric is currently too simplistic to do more than prevent extreme cases. The metric begins by tallying the number, \( n \), of image analysis
features that have similar values between the two images (i.e. that fall within a specified
distance of each other). This can be expressed with the following equation:

\[ n = \sum_i \left\lceil 0.3 - |F_i^S - F_i^P| \right\rceil \]  

(3.3)

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

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

(3.4)

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

Fitness-based tournament selection determines those genotypes that propagate to the next generation and those genotypes that participate in crossover. One-point “cut and splice” crossover is used to allow for variable length offspring. Crossover is accomplished in two stages: the first occurs at the filter level, so that the two genomes swap an integer number of filters; the second occurs at the parameter level, so that filters on either side of the cut point swap an integer number of parameters. By necessity, parameter list length is preserved for each filter. Table 3.1 shows the parameter settings used.

Mutation also occurs at two levels. Filter mutation is a wholesale change of filter (discrete values), while parameter mutation is a change in parameter values for a filter (continuous values). When filter mutation occurs, either a single filter within a genotype changes or a new filter is added. When a parameter mutation occurs, anywhere from one to all of the parameters for a single filter in a genotype are changed. The degree of this...
Table 3.1: Parameters used for the evolutionary mechanism.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Sub-Populations</td>
<td>8</td>
</tr>
<tr>
<td>Size of Sub-Populations</td>
<td>15</td>
</tr>
<tr>
<td>Crossover Rate</td>
<td>0.4</td>
</tr>
<tr>
<td>Filter Mutation Rate</td>
<td>0.03</td>
</tr>
<tr>
<td>Parameter Mutation Rate</td>
<td>0.1</td>
</tr>
<tr>
<td>Migration Rate</td>
<td>0.2</td>
</tr>
<tr>
<td>Migration Frequency</td>
<td>0.1</td>
</tr>
<tr>
<td>Tournament Selection Rate</td>
<td>0.75</td>
</tr>
<tr>
<td>Initial Genotype Length</td>
<td>2 to 4 filters</td>
</tr>
</tbody>
</table>

change, $\Delta f_i$, for each parameter, $i$, is determined by one of the following two equations chosen randomly with equal probability:

$$\Delta f_i = (1 - f_i) \cdot \text{rand} \left( 0, \frac{(|f|+1) - |\Delta f|}{|f|} \right)$$  \hspace{1cm}   (3.5)$$

$$\Delta f_i = -f_i \cdot \text{rand} \left( 0, \frac{(|f|+1) - |\Delta f|}{|f|} \right)$$  \hspace{1cm}   (3.6)$

Here, $|f|$ is the total number of parameters in the mutating filter, $|\Delta f|$ is the number of changing parameters in the mutating filter, and $\text{rand}(x, y)$ 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).

### 3.4 Methods and Results

The evaluation of artifacts is very subjective, making an evaluation of DARCI non-trivial. Furthermore, the quality of the artifacts that DARCI produces can be judged based on two distinct criteria: how well the artifacts portray the synsets dictated by a commission, and how well the artifacts demonstrate artistic skill. Depending on the synsets in question, the first criterion can be considered less subjective than the second. For example, if the synset blue, as in the color blue, were chosen, the degree to which an artifact possesses the color blue could be measured quite objectively. As less simple/concrete synsets are applied, this criterion becomes increasingly subjective; however, we argue that it will never be more subjective than a general assessment of artistic merit. For this reason, we have chosen to focus on the first criterion of quality and relegate the second criterion to an interesting side note in this paper.

Despite focusing on the first criterion of quality, we want to eventually move in the direction of artistic analysis of DARCI's artifacts. Thus, we have selected three synsets that, while dictating some expected traits within an image, also prescribe subjective features within an image. The synsets we have selected are “fiery” as in like or suggestive of fire, “happy” as in enjoying or showing or marked by joy or pleasure, and “lonely” as in lacking companions or companionship. These synsets are well represented in DARCI’s database and are distinct in meaning.

Because there is always a subjective component in determining whether an image can be described by a given adjective, the most objective way that we can evaluate such quality is through a combination of many personal opinions. For this reason, we designed a survey in which people rank DARCI’s artifacts, alongside several other artifacts, with respect to how well the images reflect particular adjectives. For this survey we selected three images...
on which to test DARCI’s rendering of the aforementioned synsets. The images we selected are shown in Figure 3.2. We chose photographs in order to accentuate the impact of the non-photorealistic rendering tools available to DARCI. The three photographs explore the light vs. dark, chromatic vs. monochromatic, and close vs. distant spectrums.

For each photograph and for each synset we commissioned DARCI to produce an image that portrays the synset; we also commissioned DARCI to produce a variation of each image that portrays all three synsets simultaneously in order to demonstrate the effect of combining synsets with disjointed meaning (this results in a total of $3 \times 3 + 3 = 12$ images).

For comparison, we collected three additional sets of 12 homologous images: a set chosen by us from a collection of images created by DARCI, a set commissioned to human artists, and a set chosen by us from a collection of randomly generated images.

For the set created by DARCI, we allowed DARCI to practice the three synsets for eight hours a piece, and then gave the system sixteen hours to complete each commission. For every commission, DARCI chose the single image with the highest fitness as the result of the commission.

In addition, for each commission, DARCI saved the top five unique images (those with the highest fitness) encountered within each sub-population, for a total of forty images. From these, we chose the single image we thought best portrayed the commission target synset(s). (We made this selection with no knowledge of DARCI’s fitness values for the 40 images, and,
in particular, we did not know which of the images DARCI ranked highest and selected as the result of its commission.) This image we selected represents a close cooperation between DARCI and DARCI’s programmers—or, looked at another way, the use of DARCI as a tool rather than as an autonomous agent.

A third set of images was created by human volunteer artists, who were restricted to a toolset similar to that used by DARCI (i.e. image filters) and were skilled with programs (e.g. Photoshop) using this toolset.

Each image in the final set was chosen from a set of 40 randomly generated images, each of which was generated using $1 - 8$ of the same filters available to DARCI. In order to ensure a reasonable image, and to provide a point of comparison between random filter generation and DARCI’s evolutionary mechanism, we chose the one image (out of 40) that we thought best portrayed the synset in question.

In summary, we acquired four images for every synset-image combination. One was DARCI’s most fit artifact (DARCI), one was our choice out of DARCI’s top artifacts (Coop), one was produced by a human (Human), and one was our choice out of randomly filtered images (Best Random). Representative examples of some of the twelve synset-image combinations can be found in Figures 3.4-3.7. In the online survey, volunteers were instructed to rank the four images for each synset-image combination according to how well they portrayed the synset(s) in question. In addition, we asked the volunteers to indicate which images they liked regardless of adjective compatibility. This additional question was added to stress to the volunteers the fact that the ranking was to be independant of personal preference for the images. We obtained a total of forty-two survey responses.

The results of this survey are encapsulated in Figure 3.3 and Table 3.2. Figure 3.3 shows the average ranking for each synset across all three images for each of the four artifact sources just summarized: DARCI, Human, Coop, and Best Random (the lower the rank, the better). Table 3.2 shows the average ranking of each of the four artifact sources across all
Figure 3.3: The average ranking for each synset across images A, B, and C for each of the four artifact sources: DARCI, Human, Coop, and Best Random. “triple” refers to the artifacts rendered with all three synsets. These results were obtained from 42 volunteers. Lower rank is better.

<table>
<thead>
<tr>
<th></th>
<th>Human</th>
<th>DARCI</th>
<th>Coop</th>
<th>Best Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Rank</td>
<td>2.4067</td>
<td>2.7282</td>
<td>2.3194</td>
<td>2.5456</td>
</tr>
</tbody>
</table>

Table 3.2: The average ranking of the four artifact sources across all image-synset combinations. These results were obtained from 42 volunteers. Lower rank is better.

synsets and images. Table 3.3 shows which pairs of datapoints in the aforementioned figures are statistically significant—such pairs are denoted with an asterisk.

3.5 Discussion

By looking at Table 3.2, we see that DARCI functioning autonomously does perform the worst of the artifact sources, but not dramatically so. Furthermore, Table 3.3 indicates that human performance was not distinguishable, in the statistical significance sense, from the performance of DARCI in cooperation with humans nor from the performance of humans choosing the best random image. These results suggest that, overall, volunteers are not strongly preferring one artifact source over another.
<table>
<thead>
<tr>
<th></th>
<th>fiery</th>
<th>happy</th>
<th>lonely</th>
<th>triple</th>
<th>all synsets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human/DARCI</td>
<td>0.501</td>
<td>*</td>
<td>1.000</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Human/Coop</td>
<td>*</td>
<td>*</td>
<td>0.155</td>
<td>*</td>
<td>0.217</td>
</tr>
<tr>
<td>Human/Best Random</td>
<td>0.286</td>
<td>*</td>
<td>0.326</td>
<td>0.339</td>
<td>0.0540</td>
</tr>
<tr>
<td>DARCI/Coop</td>
<td>*</td>
<td>*</td>
<td>0.132</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>DARCI/Best Random</td>
<td>0.691</td>
<td>*</td>
<td>0.298</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Coop/Best Random</td>
<td>*</td>
<td>1.000</td>
<td>0.640</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

* p-value < 0.01

Table 3.3: Results of $t$-Test comparing all binary combinations of artifact sources for each synset. The “all synsets” column refers to Table 3.2. The other columns refer to Figure 3.3.

When looking at performance over individual synsets (Figure 3.3), we see that a more distinct preference is given to certain artifact sources over others. But, even in these cases, the source given preference varies from synset to synset. Looking at Figure 3.3, the clearest distinction between sources is between the human and autonomous DARCI when rendering “happy” images. In this case humans clearly outperform DARCI. However, in the case of “fiery” images, DARCI performs statistically the same as humans. When in cooperation with humans, DARCI significantly outperforms solo humans in both “fiery” images and images combining all three synsets. In the case of “lonely” images, none of the artifact sources perform statistically different from one another. Volunteers prefer human creations for “happy” images and they prefer DARCI-human collaborations for both “fiery” images and images combining “fiery”, “happy”, and “lonely”.

If we look even more specifically at the individual synset-image pairs, we find that all artifact sources are top ranked for some of the pairings. Autonomous DARCI is top ranked for “fiery” image A and “lonely” image C; the best-of-random source is top ranked for “happy” image C, “lonely” image A, and “triple” image A; DARCI in cooperation with humans is ranked top for “fiery” image B, “fiery” image C, and “triple” image B; humans creating solo are ranked top for “happy” image A, “happy” image C, “lonely” image B, and “triple” image C. The rankings for the most substantial successes of each artifact source are shown in Figures 3.4-3.7.
Figure 3.4: Image A rendered “fiery”—ranked left-to-right from most to least fiery.

Figure 3.5: Image A rendered “happy”—ranked left-to-right from most to least happy.

While DARCI’s solo artifacts often rank on par with human artifacts, the best random artifacts do as well. Furthermore, these partially random artifacts are sometimes ranked better than DARCI’s. If these were totally randomly generated artifacts, then this would be an area of concern. It turns out, however, that given the number of random images from which we selected, it is fairly common to encounter at least one image that (at least to some extent) satisfies the demands of the synset in question. Taking into account Ritchie’s proposal that the proportion of high quality artifacts produced should be correlated with creativity [77], and observing DARCI’s top forty artifacts, it becomes clear that DARCI is accomplishing something better than random image generation. Figure 3.8 shows the 40 images DARCI chose to save while rendering image A as “fiery”, while, for comparison, Figure 3.9 shows the 40 random images generated for the same task. While we did not empirically determine the proportion of images in these sets that are “fiery”, it is apparent that significantly more images are “fiery” in Figure 3.8 than in Figure 3.9.
3.6 Conclusions

If we assume that the human artists commissioned to produce artifacts for this research did indeed produce renderings that portray the synsets, then we conclude that, given the same toolset, DARCI can also produce renderings that portray them. This is a compulsory assumption since by the nature of art, the only way DARCI can be evaluated as an artist, is in comparison to other (human) artists. While on the whole, at this point people tend to favor human solo works over DARCI’s solo works, the differences are not substantial or consistent enough to warrant a different conclusion. Furthermore, the collaboration between DARCI and human’s was frequently favored over human solo artifacts. This indicates the potential for DARCI to be used as a tool to augment the creative process of human artists.

Only three synsets were tested in this experiment. However, these synsets are representative of the meaning that we want DARCI to be able to incorporate into artifacts to facilitate visual communication with an audience. DARCI has sufficient data, ergo sufficient appreciation to perform similarly on many more synsets. We are currently updating DARCI
so that the system can perform commissions online while using any known synsets. This will allow us to further observe DARCI’s capacity for rendering.

In future work regarding the evaluation DARCI, we will be exploring Ritchie’s other criteria for creativity: namely novelty and typicality. In addition, we will explore the artistic side of quality, rather than the strictly pragmatic one explored in this research (i.e. the degree to which synsets were incorporated into the artifacts).
Figure 3.9: The forty randomly generated renderings of image A for the synset “fiery”.

3.7 Acknowledgments

Warm thanks to Simon Colton for providing us with Filter Feast image filters that were included in DARCI’s toolset.

This material is based upon work supported by the National Science Foundation under Grant No. IIS-0856089.
Chapter 4

Finding Creativity in an Artificial Artist¹

Abstract

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

4.1 Introduction

DARCI (Digital ARtist Communicating Intention) is a persistent computer system currently designed to automatically render images to match a list of adjectives. For example, if someone were to request it, DARCI could render an image of Van Gogh’s Starry Night so that it could be described as happy and funny. DARCI is part of our research exploring the perception of creativity in an artificial system. In designing and evaluating DARCI we have used Colton’s creative tripod as a reference for the requirements of creativity [12]. This paper illustrates how DARCI functions according to Colton’s creative tripod and how we can evaluate the creativity of DARCI and other artificial systems.

With the creative tripod, Colton posits three behaviors required for a system to be considered creative: skill, imagination, and appreciation. These behaviors correlate strongly with traditional notions of creativity including the standard psychological definition of creativity that requires both originality and functionality, and the idea widely accepted in the field of computational creativity that novelty and quality are central to the creative process [6]. Colton’s skill is the ability to produce functional, or quality, artifacts that are recognized as members of their intended domain. Imagination is the ability to produce original artifacts with the added caveat that they are meaningful. In other words, they are not the result of entirely random processes. Appreciation is the capability of the system to evaluate the artifacts it produces.

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

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

4.2 Background

There have been many systems designed to produce art. Some of the more prominent systems recognized in the budding field of computational creativity include Harold Cohen’s AARON [61] and Simon Colton’s Painting Fool [13]. Both of these systems are personified to some extent by their creators and some, if not all, of their inner workings are kept from the public. We see DARCI as a system analogous to these, although our intention is to disclose all of DARCI’s mechanisms in our research.

Fundamental to DARCI’s creation process is an evolutionary mechanism which we will describe in detail shortly. Evolutionary art, such as the images DARCI produces, has a rich background with many developments. Traditionally, evolutionary art has been an interactive affair involving the creation of a virtual space of myriad possible images for a human artist or artists to explore. This space of possible images is defined by a digital encoding, or genotype, and the rules to transform this genotype into an actual image, commonly called the phenotype. Evolutionary mechanisms begin with a population of random genotypes which are evaluated based on the qualities of their respective phenotypes. This evaluation is effected by a fitness function and determines which genotypes get to pass on their traits to future generations of the evolutionary mechanism. Since art is inherently subjective and difficult to parameterize, this evaluation is often left to human judgment. Hence a human’s aesthetic sense acts as the fitness function for most evolutionary computation used for producing art [78, 80, 83, 86].
Unfortunately, evolutionary mechanisms usually require large populations and many
generations to converge to interesting images. This can be tiresome to human evaluators,
not to mention eliminates much of the autonomy that we are interested in establishing in a
creative system. Other evolutionary systems have been developed that automatically assign
a fitness function to phenotypes. Most of these systems extract quantifiable features from the
images to evaluate them. The hope is that aesthetic images will contain certain measurable
features. Examples of features used in existing evolutionary systems include: how closely
the image color distribution matches the color distribution of highly rated Flickr images [73],
how complex the image is as measured by image compression [56], a geometric assessment of
regions within an image [36], etc.

More sophisticated approaches use a dynamic fitness function that changes based on
conditions in the evolutionary environment. For example, DiPaola and Gabora have designed
a system that examines two sets of features when assigning fitness to images: one is the
degree to which an image is similar to a target image and the other is the degree to which an
image follows quantifiable rules of aesthetics. The weight that the algorithm puts on each set
of features changes whenever evolution becomes stagnant [27]. Another example of a dynamic
fitness function is a co-evolutionary model implemented by Greenfield. In this algorithm, the
system co-evolves a population of images and a population of image filters together. The
fitness functions of each population differ, but are dependent upon the interaction between
the two populations (i.e. how the image filters affect the images). Thus, the fitness functions
fluctuate as if in an ecological setting [37].

Another approach to automating fitness evaluation that has been explored is to model
natural systems that can be used to evaluate images. For example, one could model a human’s
appreciation of images and use that model as the fitness function rather than an actual human.
This particular case was first explored by Baluja et al. using an artificial neural network to
model the aesthetic preferences of specific users [2]. To train the neural network, Beluja used
data collected from a traditional human-as-the-judge evolutionary mechanism. The images
that the user selected each generation were labeled positively as aesthetic images. The entire pixel space of each image was the input for the neural network. Beluja had minimal success noting the difficulty in using such an enormous input space for such a sparse selection of data points.

Recently, Machado et al. augmented their NEvAr system by incorporating a dynamic artificial neural network model to evaluate the novelty of the system’s creations [57]. They trained their new model to distinguish between two pools of images: those that NEvAr created and paintings by renowned artists. As suggested by Beluja, for the inputs into the neural net they used carefully selected image features rather than the entire pixel space. Machado then used the model to act as the fitness function for NEvAr’s evolutionary mechanism. Images that the system recognized as its own creations were rejected, thus the system had to evolve to change its own style. After terminating the evolutionary mechanism, the rejected creations were added to the neural network model as NEvAr’s creations. The experiment was repeated again. This process was repeated twelve times forcing NEvAr to dramatically change its style.

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

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

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

Csíkszentmihályi explained the necessity of social context to define creativity [21]. DARCI is designed to learn and create within such a context primarily through the system’s interactions with people via the internet. Ultimately, DARCI’s representation of adjectives is a reflection of those who have trained the system using DARCI’s website. As previously mentioned, traditional genetic algorithms require social interaction as well; however, these traditional algorithms involve the explicit creation of artifacts as opposed to the creation of a language—in DARCI’s case, a language for communicating adjectives with images.

Previous research has examined DARCI’s ability to produce images that communicate specific meaning [66, 68]. While touching on that topic, this paper will focus more on DARCI’s ability to produce creative images in general, as well as how we can measure that creativity.

4.3 Methods

DARCI is an elaborate system with several interdependent components. These components can be grouped into the two principal processes that allow DARCI to be perceived as creative. The first process corresponds with Colton’s appreciation behavior while the second corresponds with both the skill and imagination behaviors. Here we describe each of the components that make up these two key processes.

4.3.1 Appreciation

Colton defines appreciation in a computational system as the ability for that system to evaluate its own artifacts. DARCI’s artwork (or artifacts) comes in the form of images. In order for DARCI to appreciate art, it must first acquire some basic understanding of art. For example, in order for DARCI to appreciate an image that is dark and gloomy, DARCI must
Table 4.1: Set of features DARCI extracts from an image in order to associate images with adjectives.

<table>
<thead>
<tr>
<th>Color &amp; Light:</th>
<th>Texture:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Average red, green, and blue</td>
<td>1. Co-occurrence matrix (x4)</td>
</tr>
<tr>
<td>2. Average hue, saturation, and intensity</td>
<td>1. Maximum probability</td>
</tr>
<tr>
<td>3. Saturation and intensity contrast</td>
<td>2. First order element</td>
</tr>
<tr>
<td>4. Unique hue count (from 20 quantized hues)</td>
<td>difference moment</td>
</tr>
<tr>
<td>5. Hue contrast</td>
<td>3. First order inverse element</td>
</tr>
<tr>
<td>6. Dominant hue</td>
<td>difference moment</td>
</tr>
<tr>
<td>7. Dominant hue image percent</td>
<td>4. Entropy</td>
</tr>
<tr>
<td>Shape:</td>
<td>5. Uniformity</td>
</tr>
<tr>
<td>1. Geometric moment</td>
<td>2. Edge frequency (25x vector)</td>
</tr>
<tr>
<td>2. Eccentricity</td>
<td>3. Primitive length</td>
</tr>
<tr>
<td>3. Invariant moment (5x vector)</td>
<td>1. Short primitive emphasis</td>
</tr>
<tr>
<td>4. Legendre moment</td>
<td>2. Long primitive emphasis</td>
</tr>
<tr>
<td>7. Edge direction histogram (30 bins)</td>
<td>5. Primitive percentage</td>
</tr>
</tbody>
</table>

first understand the concepts *dark* and *gloomy*. To do this, DARCI must learn to associate images with artistic descriptions.

Before DARCI can form associations between images and descriptive words, the appropriate image features for the task must be extracted from the image. We use low-level features that can characterize the various ways that an image can be appreciated. There has been a large amount of research done in the area of image feature extraction [24, 35, 51, 53, 91, 92]. We select 102 image features from all of these areas of research. Our set of image features are broken down into the areas of color, light, texture, and shape as shown in Table 4.1.

DARCI forms an appreciation of art by making associations between image features and image descriptions. An image can be described and appreciated in many ways: by the subject of the image, by the aesthetic qualities of the image, by the emotions that the image evokes, by associations that can be made with the image, by the meanings found within the image, and possibly others. To teach DARCI how to make associations with such
descriptors, we present it with images labeled appropriately. Eventually we would like DARCI to understand images from all of these perspectives. However, for now, we have reduced descriptive labels exclusively to delineated lists of adjectives.

We use WordNet’s [30] database of adjectives to give us a large, yet finite, set of descriptive labels. Even though our potential labels are restricted, the complete set of WordNet adjectives can allow for images to be described by their emotional effects, most of their aesthetic qualities, many of their possible associations and meanings, and even, to some extent, by their subject. In WordNet, each word belongs to a synset of one or more words that share the same meaning. If a word has multiple meanings, then it can be found in multiple synsets. Our image classification labels actually consist of a unique synset identifier, rather than the adjectives themselves.

To collect training data we have created a public website for training DARCI (http://darci.cs.byu.edu). From this website, users are presented with a random image and asked to provide adjectives that describe the image. When users input a word with multiple senses, they are presented with a list of the available senses, along with the WordNet gloss, and asked to select the most appropriate one. Additionally, for each image presented to the user, we list seven adjectives that DARCI associates with the image. The user is then allowed to flag those labels that are not accurate. This creates strictly negative examples of those synsets, which will be important in the learning process.

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

Learning image to synset associations is a \textit{multi-label classification} problem [87], meaning each image can be associated with more than one synset. We cannot assume that each training image will be labeled with all the possible correct synsets. As we train the appreciation network on an image, we only train the ANNs that are explicitly labeled (as positive or negative), and ignore the other neural nets.

ANNs require a lot of training data to converge. Currently, of the 211 synsets known to DARCI, there are on average just over 33 positive data instances per synset. In order to enhance the amount of positive and negative data used to train the appreciation network, we utilize synset relationships built into WordNet in addition to statistical correlations present in our data as described in previous work [66].

4.3.2 Skill and Imagination

Because Colton’s notions of skill and imagination are very closely linked in our work, we will discuss them together. DARCI demonstrates skill by rendering images to match given lists of adjectives. DARCI demonstrates imagination by creating these artifacts in a non-random way while producing unpredictable results that still reflect the original images.

Multiple definitions of skill emphasize the fact that a skill must be learned or acquired through training. Also, because of the innate ability of evolution to yield unpredictable solutions to problems, evolutionary methods are frequently used in computational creativity experiments [33]. Thus, in order to argue for skill and imagination in DARCI’s creation process, we have implemented an evolutionary mechanism to render images so that they visually express the meaning of given synsets.

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

The genotypes that comprise each gene pool are lists of filters, and their accompanying parameters, for processing an image. Many of these filters are similar to those found in Adobe Photoshop and other image editing software. Others come from a series of 1000 filters that Colton et al. discovered using their own evolutionary mechanism [16]. 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 4.1 gives an example of a genotype and its effect on a sample image. There are a total of sixty-one traditional filters that we selected for DARCI to use and a total of thirty-one categories of filters from Filter Feast, making ninety-two filters available for each genotype. We selected traditional filters that were easily accessible, diverse, fast, and that didn’t incorporate alpha values (since our feature extraction techniques cannot yet process alpha values).

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

\[ \text{Fitness}(g) = \lambda_A A(g) + \lambda_I I(g) \]  

where \( g \) is an image artifact and \( A : G \rightarrow [0, 1] \) and \( I : G \rightarrow [0, 1] \) are two metrics: appreciation and interest. These compute a real-valued score for an image artifact (here, \( G \) represents the set of all image artifacts). \( \lambda_A + \lambda_I = 1 \), and for now, \( \lambda_A = \lambda_I = 0.5 \).
Both metrics used in the fitness function are applied to the phenotype (the image that results when each genotype is applied to a source image). The fitness of every phenotype within a generation of the evolutionary mechanism is determined using the same source image; but, the source image used from generation to generation depends upon which mode the system uses. In commission mode, the source image is the same from generation to generation, while in practice mode the source image for each generation is randomly selected from DARCI’s growing image database.

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

$$A(g) = \sum_{w \in C} \alpha_w \text{net}_w(g) \tag{4.2}$$

where $C$ is the set of synsets to be portrayed, $\text{net}_w(\cdot)$ is the output of the appreciation network for synset $w$, $\sum_w \alpha_w = 1$, and $\alpha_w = 1/|C|$ (though this can, of course, be changed to weight synsets unequally).
The interest metric $I$ penalizes phenotypes that are either too different from the source image, or are too similar. The metric begins by tallying the number, $n$, of image analysis features that have similar values between the two images (i.e. that fall within a specified distance of each other). This can be expressed with the following equation:

$$n = \sum_i \left[ 0.3 - \left| F^S_i - F^P_i \right| \right]$$

(4.3)

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

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

(4.4)

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

Fitness-based tournament selection determines those genotypes that propagate to the next generation and those genotypes that participate in crossover. One-point “cut and splice” crossover is used to allow for variable length offspring. Crossover is accomplished in two stages: the first occurs at the filter level, so that the two genomes swap an integer number of filters; the second occurs at the parameter level, so that filters on either side of the cut point swap an integer number of parameters. By necessity, parameter list length is preserved for each filter. Table 4.2 shows the parameter settings used.

Mutation rate is the probability that a mutation will occur in each genotype. Parameter mutation rate is the probability that when a mutation occurs, it is a parameter mutation; otherwise, it is a filter mutation. Filter mutation is a wholesale change of a single filter.
Number of Sub-Populations  8  
Size of Sub-Populations  15  
Crossover Rate  0.4  
Mutation Rate  0.1  
Parameter Mutation Rate  0.9  
Migration Rate  0.2  
Migration Frequency  0.1  
Tournament Selection Rate  0.75  
Initial Genotype Length  2 to 4 filters  

Table 4.2: Parameters used for the evolutionary mechanism.

(discrete values), while parameter mutation is a change in parameter values for a filter (continuous values). When a parameter mutation occurs, anywhere from one to all of the parameters (uniformly chosen) for a single filter in a genotype are changed. The degree of this change, $\Delta f_i$, for each parameter, $i$, is determined by one of the following two equations chosen randomly with equal probability:

$$\Delta f_i = (1 - f_i) \cdot \text{rand} \left( 0, \frac{(|f|+1) - |\Delta f|}{|f|} \right)$$  \hspace{1cm} (4.5)$$

$$\Delta f_i = -f_i \cdot \text{rand} \left( 0, \frac{(|f|+1) - |\Delta f|}{|f|} \right)$$  \hspace{1cm} (4.6)$$

Here, $|f|$ is the total number of parameters in the mutating filter, $|\Delta f|$ is the number of changing parameters in the mutating filter, and $\text{rand}(x, y)$ 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).

4.4 Evaluation of Creativity

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

4.4.1 Evaluation of Appreciation

DARCI’s appreciation is effectively a measure of the system’s ability to label images with
adjectives. While the process of determining what adjectives (or synsets) describe an image
is subjective, learning these associations is essentially a multi-label classification problem
that we can measure by using the following standard multi-label classification evaluation
metrics [79, 95].

- *Hamming Loss* - The average percentage of correct synsets not predicted and incorrect
  synsets predicted.
- *Precision* - The average percentage of predicted synsets that are correct.
Table 4.3: Ten-fold cross-validation results of DARCI’s appreciation network compared with a random untrained ANN. Lower numbers indicate better performance for Hamming loss, one-error, and coverage, while higher numbers are better for precision and recall. The difference between DARCI’s networks and random ANNs are statistically significant for a p-value < 0.001 across all metrics.

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

Synsets are ranked according to a function of the activation values of their respective neural net output nodes [66]. A synset is predicted when its corresponding output node has an activation value greater than or equal to the threshold of 0.5. For the above metrics, DARCI’s predictions are compared against the user labels contained in the database of images that we have collected using the aforementioned website. Thus, the correct results are the opinions of anonymous volunteers using the website to train DARCI. Table 4.3 shows the results of the evaluation metrics comparing DARCI to a random untrained neural network. These results are the average of ten-fold cross-validation where 90% of the labeled data is used to train DARCI’s neural nets, and 10% of the data is used to test the neural net outputs as explained. The metrics are calculated ten times, each time using different 90% and 10% blocks of data to train and test the neural nets.

It is important to note that these metrics are only applied to the synsets explicitly given to the images. In other words, the average image only has approximately nine synsets labeled as positive or negative (out of the 211 possible). This means we can only measure the

<table>
<thead>
<tr>
<th>Metric</th>
<th>DARCI</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hamming Loss</td>
<td>0.307</td>
<td>0.502</td>
</tr>
<tr>
<td>Precision</td>
<td>0.606</td>
<td>0.390</td>
</tr>
<tr>
<td>Recall</td>
<td>0.563</td>
<td>0.446</td>
</tr>
<tr>
<td>One-Error</td>
<td>0.297</td>
<td>0.579</td>
</tr>
<tr>
<td>Coverage</td>
<td>0.347</td>
<td>0.581</td>
</tr>
</tbody>
</table>
(a) blue, cold, reflective, beautiful, tranquil, glazed, patterned

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

Figure 4.2: Images that DARCI has interpreted. Words underneath each image are the adjectives DARCI associated with each image. Images courtesy of Shaytu Schwandes.

performance on those nine synsets because we don’t know anything about the remaining 202 synsets. Ideally, we would like a test set that has labels for all 211 synsets for each test image in order to fully evaluate how well DARCI can generalize. Despite this limitation, ten-fold cross-validation still gives us a good idea of how well the appreciation network is learning.

Figure 4.2 shows DARCI describing two images on which DARCI has not been trained, and a case can be made for describing each image the way DARCI did. These results show that DARCI is making progress in being able to appreciate images. This ability to appreciate images has been shown to be important for evaluating the artifacts DARCI produces.

4.4.2 Evaluation of Skill and Imagination

As previously mentioned, creativity is generally considered to be a function of quality and novelty. These attributes of creativity correspond to Colton’s suggested skill and imagination behaviors and are observable in the final product of creativity, the artifact itself. In order to attribute creativity to DARCI, we need to know how creative (i.e. novel and valuable) DARCI’s artifacts are. Since DARCI is learning to produce images that reflect adjectives, one measure of quality could arguably be the fidelity of the adjective representation. We have
studied this measure in previous work [66, 68]. While there is value to this metric, it clearly
does not fully reflect creativity. First of all, there is no measure of novelty. Additionally,
there is no measure of intrinsic value in the way a human critic might attribute it. In the
domain of visual art there needs to be some sense of skill and aesthetics on order to attribute
value to an artifact.

In order to evaluate quality more completely, while also evaluating novelty, we have
devised a series of six questions for human volunteers to answer about DARCI’s artifacts.
These questions were designed to both explicitly and implicitly determine how an individual
felt about a particular artifact from a creativity standpoint. We designed these questions as
five-point Likert items [54]; volunteers answered how strongly they agreed or disagreed (on a
five point scale) with a statement as it pertained to one of DARCI’s images. Following are
the six statements that we used (abbreviation in parentheses):

“I like the image.” (like)
“I think the image is novel.” (novel)
“I would use the image as a desktop wallpaper.” (wallpaper)
“Prior to this survey, I have never seen an image like this one.” (never seen)
“I think the image would be difficult to create.” (difficult)
“I think the image is creative.” (creative)

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

The images used in the online survey were randomly chosen out of 324 images created from ten commissions assigned to DARCI. Each commission consisted of the same source image, and one of ten adjectives. The adjectives we selected were: bright, cold, creepy, happy, luminous, peaceful, sad, scary, warm, and weird. These adjectives were selected because DARCI had ample training data for each and they represent a good diversity of meaning and relationships. For each commission we obtained the five best images from each of the eight sub-populations after running the commission gene pools for fifty generations each. Several of the resulting 400 images were nearly or exactly identical. These were removed in order to prevent duplication artifacts in the statistical analyses we ran on the data. A high scoring example image for each of the ten adjectives, along with the source image, is shown in Figure 4.3. Though the images were selected randomly, every ten images a volunteer completed in a single session consisted of an image from each of the ten unique adjectives.

For each of the ten adjectives represented, the average score for each Likert item and the combined average score was calculated and can be found in Table 4.4. For each Likert item, the statistical significance of the difference between the highest scoring adjective and each remaining adjective was calculated using a two-tailed t-test. Those differences with a p-value > 0.05 are indicated in the table, thus identifying which scores are statistically comparable with the top scoring adjective for each item. Figure 4.4 shows the average combined score for each of the adjectives in graph format for easy comparison. From this table and graph it is clear that the adjectives peaceful, scary, and weird resulted in the most creative images according to our survey while bright, luminous, and warm resulted in the least creative images.
Figure 4.3: Images that DARCI created from the same source image for ten different adjectives. These were images that received high scores from volunteers across all survey items. Source image courtesy of Jan Messersmith.

4.4.3 Effectiveness of Survey

In a Likert scale survey, it is important that the items are correlated with each other. Since we are interested in evaluating the creativity of DARCI’s images, the six items should all measure creativity to some degree. In order to assess the quality of the survey items as measurements of creativity, we calculated the Cronbach alpha of the survey with respect to these items [20]. Our survey received a Cronbach alpha of 0.838 indicating a high degree of consistency. In order to determine which questions were the most pertinent to the consistency of the survey, we calculated the Cronbach alpha with each question omitted. The results are shown in Table 4.5. From this we see that the most important item, in terms of consistency with the other items, was the statement: “I think the image is creative.” Removing this statement resulted in the greatest drop in alpha value. Since this item was the only item
directly asking about creativity, we conclude that the average survey results do indeed offer a valid measure creativity. The least important item was “Prior to this survey, I have never seen an image like this one.” Removing this question actually resulted in a higher alpha than had the statement remained, which indicates that this particular item may not be useful for our survey. Interestingly, it is the only item where a positive response could clearly have negative implications with respect to creativity—a volunteer may have never seen an image as awful as the one presented. Every other item omission resulted in a lower alpha value. From this we conclude that all of the items, with the possible exception of the “never seen” item, are valuable to our survey. Considering the “never seen” item also resulted in the least significant difference between adjective results (Table 4.4), in future surveys we will probably revise or remove the item all together.

The questions were mostly valid, but how consistent were the volunteers? Our results are only useful if the volunteers were able to accurately indicate how they felt about each image. To answer this question, we calculated Cronbach’s alpha with respect to each data point. Each data point corresponded to a specific volunteer and image. It was necessary to

Table 4.4: The average score of each adjective for each survey item. The item titles refer to the full statements listed in the paper. The highest scoring adjective for each item is heavily emphasized. The lightly emphasized values indicate scores that are not statistically different from the highest scoring adjective for each item. The statistical significance of score differences was calculated with a two-tailed t-test using a p-value threshold of 0.05. Only the statistical significance of the difference between the highest scoring adjective and the remaining adjectives under each item is evaluated.

<table>
<thead>
<tr>
<th></th>
<th>Like</th>
<th>Novel</th>
<th>Wallpaper</th>
<th>Never Seen</th>
<th>Difficult</th>
<th>Creative</th>
<th>All Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bright</td>
<td>2.509</td>
<td>2.246</td>
<td>1.912</td>
<td>2.316</td>
<td>2.316</td>
<td>2.702</td>
<td>2.333</td>
</tr>
<tr>
<td>Cold</td>
<td>2.697</td>
<td>2.909</td>
<td>1.939</td>
<td>3.121</td>
<td>2.758</td>
<td>3.030</td>
<td>2.742</td>
</tr>
<tr>
<td>Creepy</td>
<td>2.321</td>
<td>2.536</td>
<td>1.768</td>
<td>3.054</td>
<td>2.804</td>
<td>2.768</td>
<td>2.542</td>
</tr>
<tr>
<td>Happy</td>
<td>2.797</td>
<td>2.729</td>
<td>2.051</td>
<td>2.712</td>
<td>2.746</td>
<td>3.220</td>
<td>2.709</td>
</tr>
<tr>
<td>Luminous</td>
<td>2.596</td>
<td>2.404</td>
<td>1.842</td>
<td>2.579</td>
<td>2.439</td>
<td>2.719</td>
<td>2.430</td>
</tr>
<tr>
<td>Peaceful</td>
<td>3.086</td>
<td>2.931</td>
<td>2.431</td>
<td>2.897</td>
<td>3.052</td>
<td>3.259</td>
<td>2.943</td>
</tr>
<tr>
<td>Sad</td>
<td>2.821</td>
<td>2.643</td>
<td>2.286</td>
<td>2.536</td>
<td>2.554</td>
<td>2.804</td>
<td>2.607</td>
</tr>
<tr>
<td>Scary</td>
<td>2.966</td>
<td>3.085</td>
<td>2.169</td>
<td>2.932</td>
<td>2.881</td>
<td>3.271</td>
<td>2.884</td>
</tr>
<tr>
<td>Warm</td>
<td>2.473</td>
<td>2.364</td>
<td>1.582</td>
<td>2.691</td>
<td>2.218</td>
<td>2.527</td>
<td>2.309</td>
</tr>
<tr>
<td>Weird</td>
<td>3.068</td>
<td>3.220</td>
<td>2.288</td>
<td>3.085</td>
<td>3.000</td>
<td>3.339</td>
<td>3.000</td>
</tr>
</tbody>
</table>
Figure 4.4: The average combined score across all survey items for each adjective DARCI rendered images for. This score indicates overall creativity of images produced with the indicated adjective. Note that the scores of peaceful, scary, and weird are not statistically different from one another (see Table 4.4).

evaluate image-volunteer combinations since each image could evoke a very different response from the same person, and different volunteers could have different responses to the same image. Since there were over 300 different images, and volunteers evaluated on average just over 7 images a piece, the chance of a volunteer evaluating the same image twice was rare and so wasn’t considered.

With respect to the consistency of volunteer evaluations our survey received an alpha of 0.986. This is an extremely high score and actually raises some concerns. Cronbach’s alpha

<table>
<thead>
<tr>
<th>Omitted Item</th>
<th>Alpha Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>None Omitted</td>
<td>0.838</td>
</tr>
<tr>
<td>Like</td>
<td>0.798</td>
</tr>
<tr>
<td>Novel</td>
<td>0.785</td>
</tr>
<tr>
<td>Wallpaper</td>
<td>0.821</td>
</tr>
<tr>
<td>Never Seen</td>
<td>0.867</td>
</tr>
<tr>
<td>Difficult</td>
<td>0.808</td>
</tr>
<tr>
<td>Creative</td>
<td>0.782</td>
</tr>
</tbody>
</table>

Table 4.5: Alpha values measuring consistency of survey questions. The lower the alpha value, the more consistent the omitted item is with the rest of the items.
Table 4.6: Alpha values measuring consistency of volunteer responses to images. Results are shown for each grouping of the data with the resulting number of items, $K$.

<table>
<thead>
<tr>
<th>Grouping</th>
<th>$K$</th>
<th>Alpha Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Grouping</td>
<td>582</td>
<td>0.986</td>
</tr>
<tr>
<td>All Sessions</td>
<td>83</td>
<td>0.954</td>
</tr>
<tr>
<td>First Sessions</td>
<td>69</td>
<td>0.942</td>
</tr>
<tr>
<td>Adjectives</td>
<td>10</td>
<td>0.978</td>
</tr>
</tbody>
</table>

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

The first grouping was by session. Each session was averaged making a total of 83 groups—still high, but much smaller than 582. This yielded an alpha of 0.954. The next grouping was by session as well but omitted all repeat sessions (sessions where the volunteer indicated that they had done the survey before). This resulted in a $K$ of 69 and an alpha of 0.942. The final grouping was by adjective. All data points for images of the same adjective were averaged together yielding a very reasonable $K$ of 10. The alpha for this final grouping was 0.978. While each of these groupings eliminates some important information by averaging data points, they all yield high alpha values. A summary of these results can be found in Table 4.6. Ultimately, these high alpha scores indicate that volunteers are indeed able to consistently articulate how they feel about each image.

Since the survey is consistent with respect to questions and volunteers, we can accept the results in Table 4.4 and Figure 4.4 as accurate indications of the creativity of DARCI’s images.
4.5 Conclusions

We have presented a system called DARCI, which is designed to produce creative artifacts in the visual arts domain, and have outlined how the system functions with respect to Colton’s creative tripod [12]. Specifically, we have illustrated how the system learns and how it creates images. We have also presented a series of widely accepted metrics that can be used to evaluate DARCI’s ability to appreciate how the meanings of adjectives are expressed in images; in other words, the degree to which DARCI can identify the adjectives that describe images. Finally, we have described a survey that attempts to measure the novelty and quality of DARCI’s artifacts.

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

It is interesting to note that the highest scoring adjective is “weird”. While not usually considered an emotion, it is the only adjective of our ten with a meaning that has some obvious overlap with certain components of creativity, such as novelty. This again provides evidence that DARCI is capable of learning words and can intentionally create artifacts.
that communicate those words. As a corollary, these results also support the notion that communicating intention is an important component of creativity. Artifacts that convey the most meaning are generally considered more creative than artifacts that convey little meaning.

We have shown that our image survey can help us with the subjective task of evaluating creativity. We have also shown that the survey questions are for the most part both consistent and relevant to the creative evaluation of an artifact, whether they are implicitly or explicitly dealing with creativity. Future work will involve using this survey to compare DARCI’s artifacts with those created by human artists. Future work will also involve modifying and extending DARCI’s capabilities to more accurately model creativity on a computer. For example, DARCI’s ability to render images that communicate meaning will be extended to allow DARCI to communicate metaphors.

We have shown that our image survey can help us with the subjective task of evaluating creativity. We have also shown that the survey questions are for the most part both consistent and relevant to the creative evaluation of an artifact, whether they are implicitly or explicitly dealing with creativity. In prior work, we have compared DARCI’s artifacts with artifacts created by human artists [68]. However, future work will involve using the survey presented in this paper to compare DARCI’s artifacts with those created by human artists. Future work will also involve modifying and extending DARCI’s capabilities to more accurately model creativity on a computer. For example, DARCI’s ability to render images that communicate meaning will be extended to allow DARCI to communicate metaphors. DARCI can also be used to help artists, teachers, and psychologists better understand and pinpoint how the creative process works in people. For example, DARCI was used as the focal point of a collaborative class between computer scientists and artists [67], as well as the juror for an interactive art show designed to explore the creation of art as an optimization problem [69]. The results we have shown in this paper further support the use of DARCI in similar situations.
Abstract

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

There has been a recent push in computational creativity towards fully autonomous systems that are perceived as creative in their own right. One of the most significant problems facing modern creative systems is the level of curation that is occurring in these systems. If a system is producing dozens, hundreds, or even thousands of artifacts from which a human is choosing a single valued artifact, then is the system truly fully autonomous? Colton has argued that for a system to be perceived as creative, it must demonstrate appreciation for its own work \[12\]. A strong implication of this is that the system must be able to do its own curation by autonomously selecting an artifact for human judgment.

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

Historically, DARCI has been successful at producing such images when curation is a joint effort between DARCI and a human \[42, 68\]. In these cases, DARCI selects a number of artifacts, and a human chooses their favorite from that selection. When DARCI curates on its own, the results have been significantly less successful. This decrease in quality is to be expected and is a phenomenon Colton and Wiggins call the latent heat effect—“as the creative responsibility given to a system increases, the value of its output does not (initially) increase ...” (emphasis added) \[15\]. Since we know DARCI is capable of producing interesting
images, we are interested in increasing the value of the artifacts the system produces when curating alone, thus decreasing the latent heat effect.

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

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

5.2 Image Rendering

The image rendering subsystem uses a series of image filters to render pre-existing images which we refer to as source images. The subsystem has access to Photoshop-like filters with varying parameters. It uses a genetic algorithm to discover the configuration and parameter settings of these image filters so that candidate artifacts will reflect target adjectives without over or under-filtering the source image [68, 70]. A genetic algorithm is used because evolutionary approaches elegantly facilitate the creation of artifacts through both combination and exploration, two processes described by Boden for generating creative products [7]. Gero
has also outlined how the processes underlying evolution are ideal for producing novel and unexpected solutions, a crucial part of creativity [33]. Finally, we have shown how evolutionary algorithms approximate some aspects of the creative process in human artists [67].

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

5.2.1 Adjective Matching

The adjective matching metric is the output of a learning subsystem of DARCI called the Visuo-Linguistic Associator (VLA). The VLA is a collection of artificial neural networks (ANN) that learns to associate image features with adjectives through backpropagation training. The original VLA has been described in detail previously [66]. Here we introduce an improved VLA.

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

Training Data

Training data for DARCI is contained in a database. Each data point consists of an adjective (the label), the sentiment toward the adjective (positive or negative), the image features associated with the adjective (the image), and a time stamp. In our research, the term adjective always refers to a unique adjective synset as defined in WordNet [30]. Hence, different senses of the same word will belong to different synsets, or adjectives.
Data points are added to the database as they are submitted by volunteers using a training website [40]. Whenever the training algorithm is invoked, new relevant data points are introduced to the learner one at a time in the submitted order. The learner consists of a series of binary ANNs, one for each relevant adjective. An adjective, and any corresponding data point, is considered relevant once there are at least ten distinct positive and ten distinct negative instances of the adjective in the database. Here, distinct means occurrences of the adjective with unique sets of image features (i.e. if an adjective is used to label the same image multiple times it only counts as one occurrence). At the moment the learner is invoked, a new neural network is created for any new adjectives that have become relevant.

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

The learner’s neural networks are trained using standard back propagation with 102 image features as inputs. These image features are widely accepted global features for content based image retrieval, and most of them are available through the DISCOVIR (DIStributed COntent-based Visual Information Retrieval) system [35, 51]. A summary of the features we use can be found in Table 5.1. These features describe the color content, lighting, textures, and shape patterns found in images. Specific to the art domain, several researchers have shown that such features are useful in classifying images according to aesthetics [24], painting genre [96], and emotional semantics [92]. As many of these researchers have found color to be particularly useful in classifying images, we added four color-based features inspired
Table 5.1: Image features used to train neural networks.

<table>
<thead>
<tr>
<th>Color &amp; Light:</th>
<th>Texture:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Average red, green,</td>
<td>1. Co-occurrence matrix (x4)</td>
</tr>
<tr>
<td>and blue</td>
<td></td>
</tr>
<tr>
<td>2. Average hue,</td>
<td>1. Maximum probability</td>
</tr>
<tr>
<td>saturation, and</td>
<td></td>
</tr>
<tr>
<td>intensity</td>
<td>2. First order element</td>
</tr>
<tr>
<td>3. Saturation and</td>
<td>difference moment</td>
</tr>
<tr>
<td>intensity contrast</td>
<td></td>
</tr>
<tr>
<td>4. Unique hue count</td>
<td>3. First order inverse element</td>
</tr>
<tr>
<td>(from 20 quantized</td>
<td>difference moment</td>
</tr>
<tr>
<td>hues)</td>
<td></td>
</tr>
<tr>
<td>5. Hue contrast</td>
<td>4. Entropy</td>
</tr>
<tr>
<td>6. Dominant hue</td>
<td>5. Uniformity</td>
</tr>
<tr>
<td>7. Dominant hue image</td>
<td></td>
</tr>
<tr>
<td>percent</td>
<td></td>
</tr>
<tr>
<td>Shape:</td>
<td></td>
</tr>
<tr>
<td>1. Geometric moment</td>
<td>2. Edge frequency (25x vector)</td>
</tr>
<tr>
<td>2. Eccentricity</td>
<td></td>
</tr>
<tr>
<td>3. Invariant moment</td>
<td>3. Primitive length</td>
</tr>
<tr>
<td>(5x vector)</td>
<td></td>
</tr>
<tr>
<td>4. Legendre moment</td>
<td>1. Short primitive emphasis</td>
</tr>
<tr>
<td>5. Zernike moment</td>
<td>2. Long primitive emphasis</td>
</tr>
<tr>
<td>7. Edge direction</td>
<td>4. Primitive uniformity</td>
</tr>
<tr>
<td>histogram (30 bins)</td>
<td>5. Primitive percentage</td>
</tr>
</tbody>
</table>

by Li’s own colorfulness features [53] to those contained in DISCOVIR. In Table 5.1 these colorfulness features are “Color & Light” numbers 4-7.

When training neural networks in batch, back propagation requires many epochs of training to converge. During each epoch, all of the training data is presented to the neural network in a random order. To imitate this with incremental learning, each new data point is introduced to the appropriate neural network along with a selection of previous data points. Along with this recycled data, additional data points are predicted from the co-occurrences of adjectives with images. By including predicted data we are able to augment the limited data we do have. Similar, but less complete, approaches to augmenting training data have been successful in the past [66].

Recycling Data

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

\[
P(\text{rank}(d)) = \frac{\text{rank}(d)}{\sum_{i=0}^{\mid D_{as} \mid} i} \tag{5.1}
\]

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

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

**Predicting Data**

To augment the training data we collect from DARCI’s website, we analyze the co-occurrence of relevant adjectives to predict additional data points. Here we say that two adjectives co-occur whenever the same image is labeled with both adjectives at least once—these labels can be negative or positive. As each new data point is introduced to the learner, co-occurrence counts (distinct images) are updated for all pairings of relevant adjectives across all four combinations of sentiment. For example, as of this paper, ‘scary’ has 26 co-occurrences with ‘disturbing’ (or ‘scary’ co-occurs with ‘disturbing’ in 26 distinct images) and 0 co-occurrences with ‘not disturbing’, while ‘not scary’ has 5 co-occurrences with ‘disturbing’ and 32 co-occurrences with ‘not disturbing’.
Once the co-occurrence counts have been updated, they are used to predict $m$ positive and $m$ negative data points to augment the new data point. $m$ is calculated as $\lfloor pn \rfloor$ where $p$ is a prediction coefficient and $n$ is defined above. For this paper, $p$ is set to 0.3. These predicted data points are not added to the database.

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

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

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

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

For each data point to be predicted, a likelihood distribution from either $\Lambda_+$ or $\Lambda_-$ is selected using the above probability distributions. Then an image is selected, using a uniform distribution, from all those images with the likelihood’s label (either $a_i$ or $\neg a_i$) that are not labeled with $a$. The label for the new predicted data point is $a$, the sentiment is the sentiment of the distribution $\Lambda$, and the features are the image features of the selected image.
Informally, data points are predicted by assuming that images labeled with adjectives that frequently co-occur with a given adjective, can also be labeled with the given adjective.

**Artificial Neural Networks**

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

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

When the VLA is accessed for the adjective matching metric, the candidate artifact being evaluated is analyzed by extracting the 102 image features. These features are then presented to the appropriate neural network and the output is used as the actual metric. Thus, as Baluja and Machado et al. have done previously, we essentially build and use a model of human appreciation to guide the creation process so that we will hopefully produce images that humans can value [2, 57]. Unlike Baluja and Machado however, our model associates images with language and meaning (adjectives), an important step in building a system that communicates intention with its artifacts.

### 5.2.2 Similarity

The similarity metric borrows from the growing research on *bag-of-visual-word* models [23, 85] to analyze local features rather than global ones as we have done previously [68]. Typically,
these local features are descriptions of points in an image that are the most surprising, or said another way, the least predictable. After such an interest point is identified, it is described with a vector of features obtained by analyzing the region surrounding the point. **Visual words** are quantized local features. A dictionary of visual words is defined for a domain by extracting local interest points from a large number of representative images and then clustering them (typically with $k$-means) by their features into $k$ clusters, where $k$ is the desired dictionary size. With this dictionary, visual words can be extracted from any image by determining to which clusters the image’s local interest points belong. A bag-of-visual-words for the image can then be created by organizing the visual word counts for the image into a fixed vector. This model is analogous to the bag-of-words construct for text documents in natural language processing. These fixed vectors can then be compared to determine image similarity.

For the similarity metric used in this paper, we use the standard SURF (Speeded-Up Robust Features) detector and descriptor to extract interest points and their features from images [3]. SURF quickly identifies interest points using an approximation of the difference of Gaussians function, which will often identify corners and distinct edges within images. To describe each interest point, SURF first assigns an orientation to the interest point based on surrounding gradients. Then, relative to this orientation, SURF creates a 64 element feature vector by summing both the values and magnitudes of Haar wavelet responses in the horizontal and vertical directions for each square of a four by four grid centered on the point.

We build our visual word dictionary by extracting these SURF features from more than 2000 images taken from the database of images we’ve collected to train DARCI. The resulting interest points are then clustered into a dictionary of 1000 visual words using Elkan $k$-means [28].

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

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

(5.3)

This metric effectively measures the number of interest points that coincide between the two images by comparing the angle between vectors $A$ and $B$. In text analysis, *cosine similarity* (the parenthetical expression contained in Equation 5.3) is typically used to compare the similarity of documents. With this metric, as the sparseness of vectors increases, the similarity between arbitrary vectors approaches 0. In our case, as vectors are quite sparse, artifacts that are even slightly different from the source would have low scores using this measure. Nevertheless, creating renderings that are very similar to the source image is trivial as it requires simply using fewer and less severe filters. Thus, despite encountering low scores from only small differences, the genetic algorithm would be able to easily converge to near perfect or even perfect scores. This interplay between a harsh similarity metric and relative ease of convergence would place too much weight on the similarity metric. In fact, auxiliary experiments have shown that when using cosine similarity, the adjective matching metric is almost ignored in artifact production.

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

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

The two conflicting metrics result in a process that is arguably transformational in nature, at least to a limited degree. Boden describes transformational creativity as that which transforms the conceptual space of a domain [6]. While the space of possible artifacts cannot change (the filters available for rendering images do not change), the evaluation of the artifacts does change through the interplay of the two metrics. This interplay occurs organically in the minimum fitness function by forcing the system to emphasize the metric that it is struggling to optimize at any given epoch during the evolutionary algorithm. The interplay of divergent metrics occurs more mechanically in the alternate and converge fitness functions by scheduling the emphasis; however, the sudden shift in metric could result in more unexpected results, a criterion of creativity emphasized by Maher [59, 60]. The scheduled approaches were inspired by Dipaola and Gabora’s work with “Evolving Darwin’s Gaze”, an installation that also evolves images under two shifting criteria [27]. Their criteria are a pixel matching metric comparing artifacts to a specific portrait of Charles Darwin, and an artistic heuristic. We anticipate that our less restrictive metrics will ultimately allow for even more surprise and variation in artifacts, while also communicating meaning (adjectives).

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

The five adjectives used were ‘happy’, ‘sad’, ‘fiery’, ‘wet’, and ‘peaceful’. These were chosen because they were well represented in our adjective matching training data and because they depict a range of distinct meanings and emotional valence. The three source images (referred to as images A, B, and C) are shown in Figure 5.1 with their corresponding resolutions.

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

5.4 Results

In this section we will discuss DARCI’s artifact selection for each experiment. While all interpretations of the images themselves are clearly subjective, we attempt to be conservative and consistent in our observations. We will discuss the artifacts in terms of the objectives of the image rendering subsystem: to depict the source image and adjective together in an
Figure 5.1: The three source images used in all experiments. Images A and C have resolutions of 1600x1200. Image B has a resolution of 1920x1200.

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

This definition of interesting is derived from two commonly proposed requirements for creativity applied to the specific goal of DARCI’s image rendering subsystem. These two requirements are, as defined by the American Psychological Association, functionality and originality; or, as Boden described them for the domain of computation, quality and novelty [6]. Since the purpose of the image renderer is to alter a source image, elimination of the source image would not be functional. Ritchie describes a related requirement that is also applicable here—that of typicality [77]. Ritchie defines typicality as the extent to which an artifact is an example of its intended class. In our case this would be a rendering of a source image as opposed to an entirely new image. The second requirement, novelty, requires that the image renderer produce renderings that are distinctive. Thus, minor or no changes to a source image would clearly suggest a failure at novelty. In an attempt to reduce the amount of subjectivity in our analysis, DARCI’s artifacts are either interesting by this definition or not. There is no attempt to rate the degree of interest.

In addition to being interesting, DARCI’s artifacts must match the intended adjective. In order to be as objective as possible, we will compare DARCI’s artifacts to images from the
VLA training data for each given adjective. These images are representative of the types of images one would find if searching google images for a specific adjective. Examples of these images can be found in Figures 5.2-5.6. Since DARCI is rendering, as opposed to composing, and due to the limitations of DARCI’s image analysis features (and indeed the limitations of the entire field of computer vision), we will be looking for similarities in color, light, and texture as opposed to similar object content.

The ‘sad’ training images (Figure 5.2) tend to be desaturated, even black and white, and/or dark with an emphasis on dull colors. The ‘happy’ training images (Figure 5.3) trend towards bright and colorful, often containing a full spectrum of colors. The ‘fiery’ training images (Figure 5.4) usually have distinct flame textures, are bright, and most are monochromatic—typically orange. The ‘wet’ training images (Figure 5.5) consist of cool colors, usually blue, and have frequent specular highlights and/or wavy patterns. Finally, the ‘peaceful’ training images (Figure 5.6) contain a variety of soft or pastel colors with a lot of smooth textures.
Ideally, the most fit artifact discovered by the genetic algorithm should be the one that best satisfies the objectives for object rendering outlined above. Thus, for most of the fitness functions, we used this method of selection. However, we anticipated that for two of the fitness functions, *alternate* and *converge* this would not be an appropriate approach. The reason for this is that both of these fitness functions only use one metric at a time, meaning that the most fit artifact discovered could only have been optimized for a single metric. The expected result would be the same as a selection from one of the control fitness functions—not an ideal balance of metrics.

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

The most fit artifact discovered for each source image in the similarity control experiments is shown in Figure 5.7. The most fit artifact discovered in each of the other experiments is shown in Figures 5.8-5.12.

First looking at the similarity results (Figure 5.7), we see that with the exception of image A, DARCI did not select nearly identical images as we might have expected. This illustrates the effect of not scaling the source images. The chosen artifacts actually had slightly higher fitness scores than the strictly scaled down source images demonstrated earlier. For comparison, the fitness score of each of these artifacts is, for artifacts produced from images A, B, and C respectively: 0.836, 0.762, and 0.860 with an average score of 0.820. That being said, these artifacts are still quite close to the source images, and any resemblances to any of the specified adjectives are obviously happenstance.

For the average fitness function, arguably all three of the ‘happy’ images convey their adjective by applying bright colored filters (Figure 5.8). All three of the ‘sad’ images are made
more sad by converting them to dark black and white images (Figure 5.9). Two out of the three ‘fiery’ images are fiery by primarily coloring with oranges and reds (Figure 5.10). Image B also looks bright and molten in texture, and some of the buildings in the background of image C almost look on fire. All three ‘wet’ images are debatably wet, mostly by implementing blue filters (Figure 5.11). Although, Image B actually looks like it is being viewed through a window soaked during a downpour. None of the ‘peaceful’ images look any more peaceful than their sources; and very little if anything has changed (Figure 5.12). With the odd exception of the ‘peaceful’ images, average does quite well at conveying adjectives; however, most of the images don’t use much more than simple color filters to do so. In our estimation, for the average artifacts, ‘happy’ B and C, ‘fiery’ B and C, and ‘wet’ B satisfy the objectives for object rendering as outlined earlier.

For the minimum fitness function, two of the ‘happy’ images, A and C, are made happy by incorporating many bright colors. Image A looks kaleidoscopic and image C has some rainbow effects. Image B seems out of place, though close inspection will reveal that it may have received a high fitness because of many bright colors as well. While perhaps
difficult to notice at first, both image A and B maintain the presence of the source image. All of the ‘sad’ images are quite dark, suggesting sadness. Image A and C may look like they have eliminated the source images, but the vague shape of the fish is visible within the squiggles of image A, and close inspection of image C will reveal many of the city lights behind the heavy distortion. The three ‘fiery’ images could be considered ‘fiery’. Image A literally looks on fire and image C looks molten. All three ‘wet’ images appear wet; as with average, this is primarily accomplished by making the images blue. Image B does look like

Figure 5.7: The most fit artifacts for each indicated source image discovered using the similarity fitness function.
the image is now reflected off of a lake, and image C is a bit bleary and wavy giving it ever so slightly the look of being underwater. With the exception of image A, the ‘peaceful’ images aren’t even recognizable, nor do they look peaceful in the way ‘peaceful’ is reflected in the training images. We’re beginning to get a sense of how DARCI interprets ‘peaceful’ though. In our estimation, of the minimum images, ‘happy’ A and C, all ‘sad’ and ‘fiery’ images, and ‘wet’ B and C satisfy the objectives for object rendering. While ‘happy’ B and ‘peaceful’ A are interesting representations of the source image, they do not convey the adjective properly.

In the case of the adjective fitness function, we see that with three exceptions (‘happy’ A, ‘sad’ A, and peaceful ‘C’), the source image is undetectable. ‘Happy’ A and ‘sad’ A do fit their adjectives, but ‘peaceful’ C does not. Interestingly, in our estimation adjective does not
Figure 5.9: The most fit artifacts for each indicated source image and fitness function for the adjective ‘sad’.

depict the given adjectives as well as average or minimum. This can be attributed in part to the system exploiting the VLA’s neural networks with extreme and unnatural image features.

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

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

5.4.2 Selection After Last Shift

As indicated earlier, the alternate and converge fitness functions need a different selection method than that used above. As suspected, using most-fit selection resulted in artifacts that were either similar to those in Figure 5.7 or completely abstract like the images produced with adjective. The assumption with alternate and converge is that even though only a single metric is in effect at each generation, the genetic algorithm will not be able to converge
to either because of constant shifts in the metric, and will instead find an interesting and unexpected solution.

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

The results of these experiments are in Figures 5.13 to 5.15. In the interest of space, we do curate these images by only showing those artifacts that are neither over nor under-filtered (i.e. interesting) based on observations similar to those made for the earlier experiments. In the case of alternate-similarity, there were no artifacts produced that weren’t under-filtered. Most had tinting or small distortions, but none were interesting.

Figure 5.13 shows interesting artifacts that were selected with alternate-adjective. This particular fitness function and selection criteria yielded the most numerous interesting
artifacts of the four configurations. In this case, all but one of the not-shown artifacts were too abstract. Of the remaining interesting artifacts, all but the unusual ‘peaceful’ images arguably convey the intended adjectives.

Next, Figure 5.14 shows interesting artifacts selected with converge-adjective. Most of the other artifacts selected obfuscated the source image too much. Here, with the exception of ‘fiery’ A and perhaps ‘fiery’ B, the images convey the intended adjectives.

Finally, Figure 5.15 shows the interesting artifacts selected with converge-similarity. While the images shown are adequately interesting, we don’t consider them as distinguished as those in the previous two examples. All of the other artifacts were too similar to the source image to warrant display. All of the displayed artifacts do convey the given adjectives.

5.4.3 Filter Sequence Length

Functionally, much of the quality of an artifact can be attributed to the length of the artifact’s genotype. The genotype is the “genetic” encoding of the artifact, and in the image rendering
subsystem is a sequence of image filters. The more filters used to render a source image, the more likely the artifact will become abstract. The fewer filters used, the more likely the artifact will not deviate from the source image. Figure 5.16 shows the average genotype length (in number of filters) for each fitness function explored in this paper over the 100 epochs of evolution. The top performing fitness functions show a comfortable balance between too many and too few filters. *Minimum* does this the best.

### 5.5 Conclusions

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

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

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

We are confident that the improvements made to the image rendering subsystem in this paper will significantly decrease the latent heat effect in DARCI. We intend to test this theory in the future by conducting a thorough online survey comparing this improved version of DARCI to other versions, and perhaps even to humans. To further improve the image rendering subsystem described in this paper, we also intend to pursue more adaptable variations of the metrics outlined here. Metrics that will adapt their output in response to the features of other metrics.
Chapter 6

Accounting for Bias in the Evaluation of Creative Computational Systems: An Assessment of DARCI

Abstract

Recent investigations into the assessment and evaluation of “creative” systems in the field of computational creativity have disclosed several problems common to research within the field. We perform a practical evaluation of the latest iteration of the creative system, DARCI (Digital ARtist Communicating Intention), attempting to address some of these problems using a specially designed, but generalizable, online human survey. Of note, we address the complications of evaluator bias that are present in all assessments of creativity. Using our evaluation, we show that within its narrow domain, DARCI is able to produce artifacts that are arguably at least as creative as human counterparts. Furthermore, these artifacts tend to be more surprising and perceived as more difficult to produce than those created by human artists.

\footnote{David Norton, Derrall Heath, and Dan Ventura. Accounting for bias in the evaluation of creative computational systems: an assessment of DARCI. Reserved for submission to Proceedings of the 6th International Conference on Computational Creativity, 2015.}
6.1 Introduction

Recent investigations into the assessment of “creative” systems in the field of computational creativity have disclosed several problems common to research within the field. The first problem is properly focusing assessments to the intended scope of a given creative system: how much should an evaluation focus on the artifacts themselves, *weak Computational Creativity*, and how much should it focus on the processes involved in creating the artifacts, *strong Computational Creativity* [1]? The second problem is determining measurable assessment criteria that can be used to determine if one version of a creative system is an improvement over another version, or to compare two different creative systems [17]. The third problem is empirically grounding the ambiguous terminology that is commonly used to describe and assess creative systems [10]. The fourth problem is picking, or designing, the best methodology to actually carry out the assessment of a given system [49]. The fifth problem, and one that is not addressed in detail by researchers in the field, is compensating for the effects of biases introduced by human evaluators when making a final assessment of a given system.

While the researchers divulging these issues have presented tantalizing theoretical solutions, few have implemented practical solutions (a noted exception is Jordanous who performs an actual meta-evaluation of existing evaluation methodologies [49]). In practice, as each of the researchers have noted, there is no straightforward solution to any of these problems. Here we perform a practical evaluation of the latest iteration of the creative system, DARCI (Digital ARtist Communicating Intention), attempting to address some of these problems using a specially designed, but generalizable, online human survey. Of note, we address the complication of biases introduced by human evaluators that are present in all current assessments of creativity. Using our evaluation, we show that within its narrow domain, DARCI is able to produce artifacts that are arguably at least as creative as human counterparts. Furthermore, these artifacts tend to be more surprising and perceived as more difficult to produce than those created by human artists.
We begin by recounting some of the issues surrounding creativity assessment in computational creativity, particularly when using human surveys, along with a mention of our proposed solutions. We then briefly describe the most recent incarnation of DARCI and its process for creating artifacts. Next, we describe how artifacts were commissioned for our experiments using DARCI and human volunteer artists, and then describe the survey in detail. Finally, we present the survey results and our subsequent analysis of DARCI.

### 6.2 Background

There has been some reticence in the community towards conducting human surveys as a means of evaluation. Bown notes that human surveys often have wide variance making them difficult to incorporate into established models of creativity [10]. In a study comparing several methods of evaluation, Jordanous concludes that human surveys were the least correct of the methods she explored [49]. She suggests that this was because participants, unsure of the definition of creativity, evaluated systems based on other factors. However, anonymous online surveys can quickly gather many responses from individuals outside of the computational creativity community. Having this outside opinion is valuable as it reduces biases that those within the community inevitably bring to assessments. We evaluate DARCI through such a survey, but, in order to reduce participant confusion and response variance, design it so that participants are asked to evaluate a variety of explicitly defined artifact qualities (that correspond to requirements of creativity) rather than being asked to directly evaluate the system’s overall creativity.

Bown stresses the inadequacy of human surveys as empirically grounding assessments since we don’t have an understanding of what the human responses mean [10]. In order to gain that understanding on some level, we develop a standard by which to judge the various artifact qualities that we measure. The standard is created by having survey participants assess human artifacts (the standard) in addition to DARCI’s.
In order to evaluate a creative system from a strong computational creativity standpoint, Colton et al. argue that the process by which a system produces artifacts must be evaluated [17]. While our survey questions do focus on the artifacts themselves, some are designed to glean opinions about DARCI’s creative process. However, in order for survey takers to evaluate this process, the survey cannot be blind. Participants in the survey will know that they are evaluating an artificial system, and bring with that knowledge unwanted biases. These biases may be negative if the viewer feels that art is an inherently human affair that automatically renders a computer’s efforts invalid. Or, they may be positive if the viewer feels that the computer has an unfair disadvantage and should thus be graded on a curve. Another possibility of positive bias is that the viewer is familiar with computational creativity, or even DARCI itself, and wants the study to succeed.

In order to evaluate DARCI’s creative process while taking into consideration the effects of evaluator bias, we design the survey to detect the level of human/computer bias in each survey taker. Then we use this information to determine the effects of survey taker bias and adjust our conclusions from the survey accordingly.

6.3 DARCI and Artifact Creation

DARCI (Digital ARtist Communicating Intention) is a creative system that we are developing to explore the bounds of computational creativity within the domain of visual art. DARCI is composed of several subsystems, each with its own creative potential, and each designed to perform an integral step of image creation from conception of an idea, to design, to various phases of implementation, to curation. The most complete subsystem, and the one that is the focus of this paper, is called the image renderer. The image renderer uses a genetic algorithm to discover a sequence of image filters that will render an image composition (produced by another subsystem) so that it will reflect a given description (selected from yet another subsystem). In this paper, we are only interested in assessing the creativity of DARCI’s
Figure 6.1: The process for building a pool of candidate artifacts. Two artificial neural networks (ANN), represented as rectangles, are trained to act together as the fitness function for a genetic algorithm which in turn creates candidate artifacts that are added to the pool. The novelty ANN uses the pool of candidates as training data for the next iteration of the genetic algorithm. This cycle is repeated for a number of epochs. The rounded rectangles indicate training data either from a dataset, or from the pool of candidate artifacts itself.

image renderer. When referring to DARCI throughout this paper, we will be specifically referring to the image rendering subsystem and not the entire in-progress system.

DARCI (the image renderer) is designed to produce a rendering for a given source image that reflects a given adjective(s) in an interesting way. As detailed in previous research, by interesting we mean that the rendering contains noticeable filtering without removing all trace of the source image [72]. In other words, the rendering is different enough from the source image that it satisfies the creativity requirement of originality while similar enough to the source image that it satisfies the creativity requirement of functionality.

To produce its artifact, DARCI first builds a pool of candidate artifacts from which to select the final rendering. The pool of candidates is created to contain a diverse range of renderings that reflect the given adjective. Figure 6.1 illustrates how DARCI’s image renderer builds this pool of artifacts. Once these candidates have been created, DARCI uses a heuristic to rank them and then selects the top ranked candidate as the final artifact.
6.3.1 Adjective ANN

The image renderer begins by training a binary artificial neural network (ANN) for the given adjective. This neural network, called here the *adjective ANN*, is trained to associate 51 carefully selected image features with the adjective using standard backpropagation and the *DARCI* dataset for training data. The 51 image features describe a variety of image qualities including color, lighting, texture, and local interest points, and were chosen from a larger set of 198 features using forward feature selection as described by Norton et al. [71]. Many of these image features are the result of psychological studies analyzing the connection between color and various affective words [58, 74, 92]. Others summarize local interest point data that is typically reserved for object detection in images [71]. Still other features come from a publicly available\(^2\) set of widely accepted global image features [51].

The adjective ANN is trained using the *DARCI* dataset. This dataset contains over 2000 images hand labeled by volunteers with adjectives from WordNet [30]. Volunteers are allowed to label images with both positive and negative uses of adjectives, eliminating the need to assume implicit negativity for training. As we have done in the past, we augment the DARCI dataset by adding negative datapoints for each positive datapoint by using related concepts and antonym relationships in WordNet [71].

6.3.2 Genetic Algorithm

Once the adjective ANN is trained, the image renderer uses a genetic algorithm to discover the configuration and parameter settings of Photoshop-like image filters that will render the source image to reflect the given adjective. Candidate filter sequences are evaluated by applying them to the source image and then using the resulting image as input to the adjective ANN. The output of the adjective ANN is the fitness score. To increase the variety of renderings discovered by the genetic algorithm, speciation is introduced by including

\(^2\)http://appsrv.cse.cuhk.edu.hk/~miplab/discovir/
sub-populations. In addition to the traditional mutation and crossover operations that are routine in genetic algorithms, migration between sub-populations can also occur.

After a number of generations of evolution, in our case 100, the renderings corresponding to the ten highest scoring filter sequences discovered per sub-population are returned. In these experiments, we use six sub-populations yielding 60 images. These select images are ordered by fitness, then added to the pool of candidate artifacts one at a time beginning with the most fit image. Images are only added to the candidate artifacts if they are determined to be sufficiently unique. In order to determine which artifacts are redundant, the system compares the 51-element feature vector \( A \) of each potential candidate with the feature vector \( B \) for each existing candidate by calculating the normalized cosine similarity of the two vectors, per Equations 6.1 and 6.2. Equation 6.1 calculates the cosine similarity of two vectors and Equation 6.2 normalizes the cosine similarity to a value between 0 and 1. If the similarity is greater than some threshold, the potential candidate is considered redundant and not added to the candidate pool. For our experiments, based on preliminary observations, we set this threshold to 0.95.

\[
similarity = \frac{A \cdot B}{\|A\|\|B\|} \quad \text{(6.1)}
\]

\[
|\text{similarity}| = \frac{1 + similarity}{2} \quad \text{(6.2)}
\]

6.3.3 Novelty ANN

Once the candidate artifacts have been selected, the genetic algorithm begins again. This time, a neural network we call the novelty ANN is trained. To train this neural net, the same positive datapoints used to train the adjective ANN are used, but the negative datapoints are the pool of candidate artifacts. Thus, the novelty ANN learns to distinguish images that are created by DARCI from those that are not. This process is similar to the process employed by Machado et al. in training NEvAr to create novel images [57].
A new genetic algorithm is initialized using the combined output of the novelty ANN and the adjective ANN as the fitness function. To combine the output of the two neural nets, the system selects the minimum output of the two classifiers as described by Norton et al. [72]. The genetic algorithm performs 100 generations of evolution using the new fitness function. This forces DARCI to produce images that reflect the given adjective and are distinct from the images produced earlier. As before, the most fit artifacts are added to the pool of candidate artifacts provided they are not redundant.

This process is repeated for several epochs, each epoch training a new novelty ANN and thus adding increasingly varied images to the pool of candidate artifacts as the system attempts to optimize renderings to the ever-changing fitness function. For our experiments, we perform a total of 8 epochs including the initial novelty-ANN-free 0th epoch. Figures 6.2 and 6.3 illustrate how candidate artifacts vary from epoch to epoch during one experiment with the adjective “cold” using Figure 6.4 as the source image.

6.3.4 Candidate Artifact Curation

Once the candidate pool has been created, DARCI must select the final rendering to present as the finished product—in other words, curate the candidates. Curating the candidates consists of two phases of ranking and selecting from the candidate pool. In the first phase, DARCI ranks the candidates by similarity to the source image and selects the top 10% for the second phase (see Figures 6.5(a) - 6.5(c)). For certain epochs, many of the candidate artifacts obscure the source image too much. This phase of curation increases the chance that the final rendering will make noticeable use of the source image.

During curation, similarity to the source image is calculated differently than when detecting redundant artifacts. This is because here we are more interested in preserving specific objects within the source image than its coloration. Color usage has been shown to correlate highly with the emotional content of images [53, 70, 92], and we would actually like the color content of the image to change in order to match the given adjective while keeping
Figure 6.2: Sample artifacts from epochs 0 through 3 of the candidate building process outlined in Figure 6.1 for the adjective “cold” and source image in Figure 6.4. Note that since the candidate pool is empty during epoch 0, the novelty ANN is not used in the genetic algorithm’s fitness function for this epoch.
Figure 6.3: Sample artifacts from epochs 4 through 7 of the candidate building process outlined in Figure 6.1 for the adjective “cold” and source image in Figure 6.4.
major objects within the source image recognizable. More than half of the image features used in detecting redundant artifacts are explicitly color features, while no color features are used during phase one of curation.

To calculate similarity for curation, first a 1000 element histogram of visual words is extracted from the source image and each candidate artifact. Visual words are quantized local image features commonly used in content based image retrieval approaches [84] where a visual word histogram functions like a bag-of-words model used in natural language processing [85]. The similarity between two images is calculated by taking the cosine similarity of the images’ visual word histograms per Equation 6.1.

In the second phase of ranking, DARCI ranks those images selected from the first phase by their association with the given adjective using the adjective ANN (see Figures 6.5(d) - 6.5(f)). The highest ranked image is then selected as the final artifact. This second phase occurs after over-filtered images have been removed in order to increase the chance that the final artifact will reflect the given adjective. This phase of ranking also reduces the possibility that the final image will be under-filtered. Figure 6.6 shows the candidates for one experiment that were selected from phase one and then ranked in phase two.
Figure 6.5: The curation process for selecting a final artifact from the pool of candidates (represented by a colored bar) with accompanying example images. (a) Each artifact in the pool of candidates is assigned a score of similarity to the source image (in this case Figure 6.4). (b) The candidates are ranked by this score. Here the score is depicted by the bar’s color gradient. (c) The top 10% of ranked artifacts are chosen for the next phase of curation. (d) The remaining artifacts are given a score of association with the given adjective (in this case “cold”). (e) The artifacts are ranked by the new score. (f) The “coldest” image is selected as the final artifact to be returned by DARCI.

6.4 Commissions

For our experiments, we commissioned DARCI and four human volunteers to produce renderings of the photograph in Figure 6.4 that depict it as either “cold”, “eerie”, or “violent”. These adjective were chosen because DARCI is able to associate them with images effectively [71], they are affective, and they haven’t been used extensively in previous studies involving DARCI. In order to keep the rendering tools available to DARCI and to the human artists as similar as possible, human artists were restricted to a subset of tools found in software packages used for photo manipulation. Specifically, human artists were given the following instructions:

You will be given a source photograph and an adjective. Using Gimp, Photoshop, or a similar program, alter the source photograph so that it reflects the adjective. You may only use Filters and Color Tools in Gimp, or Filters and Adjustments in Photoshop, to make alterations.
Figure 6.6: Artifact candidates selected from phase one of curation and then ranked at phase two of curation. These results were obtained from one experiment given the adjective “cold” and the source image shown in Figure 6.4. The artifacts are ordered from top left to bottom right where the top left image is the final artifact returned by DARCI.

You may not use selection tools to apply filters to selected regions. You may not use any brush-style tools, i.e. liquefy in Photoshop. You may not incorporate any other images. You may not alter the resolution or aspect ratio. There are no other restrictions.

They were then given the examples of human produced renderings from a previous study.
All four human volunteer artists have experience working with photo manipulation software. The 12 images they produced for our study are in Figure 6.7.

We commissioned DARCI seven times for each of the three adjectives. Each commission produced one artifact as outlined in the previous section. In order to match the number of human commissions for our experiments, we selected four of the seven artifacts DARCI produced for each adjective. We made the final decision to ensure varied artifacts and to eliminate potential outliers. Figure 6.8 shows all of the chosen artifacts produced by DARCI. Figure 6.9 shows those artifacts produced by DARCI that we did not include in our study.

6.4.1 Online Survey

Prior to taking the survey, all participants were informed that the survey was to help with research regarding DARCI, “a computer program we created”. Furthermore, the survey began by informing volunteers that “the results will be used in research exploring creativity in computational systems”.

The survey was separated into two parts. The first part was designed to detect any pre-existing human/computer bias in the survey taker as well as any bias the survey taker may have towards our research in particular (given the survey preface and disclosure of our system). The second part was designed to gather survey takers’ opinions about the renderings created by DARCI and the human artists.

Part 1

In the first part of the survey, volunteers were presented with 15 pairs of images. All images were created by first applying random filters from DARCI’s toolset to random source images, and then hand picking intriguing and abstract creations from the thousands of random images. To the extent possible, images that we deemed similar were paired with each other. These image pairs were presented to volunteers in a random order and with random labels. The labels indicated that one of the images was created by a human, and the other was created
Figure 6.7: Renderings of Figure 6.4 created by four human artists. The renderings were created to depict, from left to right, the adjectives “cold”, “eerie”, and “violent”. 
Figure 6.8: Renderings of Figure 6.4 created by DARCI and selected for the study. The renderings were created to depict, from left to right, the adjectives “cold”, “eerie”, and “violent”.

(a) DARCI 1

(b) DARCI 2

(c) DARCI 3

(d) DARCI 4
Figure 6.9: Renderings of Figure 6.4 created by DARCI and not selected for the study. The renderings were created to depict, from left to right, the adjectives “cold”, “eerie”, and “violent”.

by a computer program. For each pair, the volunteers were asked which they thought was the better image, and given only 10 seconds to respond.

Before beginning this portion of the survey, participants were given the following instruction:

In this part, you will be presented with two images produced using computer software. One was created by a human using digital tools; the other by a computer program using similar digital tools. Please indicate which image you think is better.
You will only have 10 seconds to make your decision, though you can advance to the next item at any time.

Since the images were randomly labeled as “human” or “computer”, unbiased volunteers should pick the “human” and “computer” options approximately equally.

Part 2

In the second part of the survey, where opinions about DARCI’s and the human artists’ images were collected, all volunteers were randomly assigned to one of three experiments: blind, basic, and detailed. The experiments were identical except for the amount of information that was presented to each volunteer. In all three experiments, volunteers were told in detail the task that they would be evaluating, namely the rendering of a source image to reflect a given adjective. This instruction was presented as follows:

In this part, you will be presented with a total of seven images. You will be asked to indicate your impressions of each image.

Each image was created by either a human artist or a computer program called DARCI (Digital ARtist Communicating Intention). The images were created using digital tools to modify a specific source photograph so that it reflected a given word.

In the blind experiment, volunteers were never given the name of DARCI (it was obfuscated from the above instruction) and were not told which images were produced by DARCI and which were produced by a human artist. In the basic experiment, volunteers were told the name of DARCI and whether each image was produced by DARCI or a human. In the detailed experiment, volunteers were not only told which images were created by DARCI, they were also given a detailed (for the layman) description of how DARCI produces its images. This description was followed by a simple one question quiz to assess comprehension.

Aside from the noted differences, the three experiments proceeded in the same manner. Survey takers were presented with the source photograph (Figure 6.4), noted as such, and
then six randomized images: one from DARCI and one from a human artist for each of the three adjectives (“cold”, “eerie”, and “violent”). Only six of the twenty-four possible images were presented to reduce fatigue. Volunteers were required to evaluate each image by indicating how strongly they agreed or disagreed with a series of 7-point Likert items [54]. To assist with these items, volunteers were allowed to view the source photograph at any time.

For the source image, the Likert items were (abbreviation in parentheses for convenience when referring to items):

“I like this image.” (like)

“This image is cold.” (cold)

“This image is eerie.” (eerie)

“This image is violent.” (violent)

For all other images, the Likert items were (adjective taking the place of the appropriate adjective):

“I like this image.” (like)

“This image is adjective.” (adjective)

“This image is a surprising modification of the source photograph.” (surprising)

“This image would be difficult to create from the source photograph.” (difficult)

“This image makes good use of the source photograph.” (use)

The items were chosen to succinctly capture certain qualities required for attributing creativity to a system via the artifacts it produces, and to a small extent, its creative process. Norton et al. showed that a similar set of Likert items were statistically consistent and correlated with participants’ opinions of creativity as a whole [70].

Researchers in computational creativity have identified several attributes necessary to attribute creativity or, as Colton has stated, not attribute un-creativity to a system. These attributes include Colton’s creative tripod (appreciation, imagination, and skill) [12],
Ritchie’s 18 criteria defined by functions of quality, novelty, and typicality [77], Jordanous’ 14 components of creativity [47], and the American Psychological Association’s functionality and originality attributes.

Many of these attributes relate to the Likert items in the survey. The like item relates to the attributes of skill, quality, functionality, and Jordanous’ ‘domain competence’ and ‘value’ components. Adjective relates to the attributes of functionality and Jordanous’ ‘intention and emotional involvement’ and ‘social interaction and communication’ (particularly in the detailed experiments) components. Surprising relates to the attributes of novelty, originality, and Jordanous’ ‘originality’ and ‘value’ components. Difficult relates to the attributes of skill and Jordanous’ ‘domain competence’ component, and emphasizes the creation process. Finally, use relates to the attributes of functionality, skill, and quality. Since DARCI produces artifacts, all of the Likert items relate to Jordanous’ ‘generation of results’ component, and for the detailed experiment where the creative process is disclosed, all of the items relate to Jordanous’ ‘progression and development’, ‘thinking and evaluation’, and ‘variety, divergence, and experimentation’ components.

6.5 Results

After removing results from volunteers who indicated that they had either taken the survey before or viewed someone else taking the survey, 284 completed surveys remained. An additional 46 surveys in various stages of completion were collected and included in calculating applicable results. A total of 100 volunteers were assigned to the blind experiments, 111 to the basic experiment, and 106 to the detailed experiment.

For purposes of evaluation, results from volunteers who failed the comprehension question were removed from the detailed results and added to the basic results. This was 68 of the 106 volunteers assigned to the detailed experiment. We also combined the detailed and basic experimental results into an single group that contained the results of all volunteers aware of the artist behind each artifact—the informed group.
6.5.1 Bias

A volunteer’s bias was calculated by subtracting the number of images they preferred labeled with “computer” from the number of images they preferred labeled with “human” in the first part of the survey. Thus, a positive score indicates a human bias while a negative score indicates a computer bias. Since the images were randomly labeled, the average bias of all test takers should have been close to 0 if there was no bias. However, the average bias of all test takers was 0.901 with a standard error of 0.185, indicating a small but significant bias either towards humans or against DARCI in particular.

When analyzing results from the second part of the survey, we averaged the scores (between 1 and 7) for each Likert item across all artifacts produced by either humans or DARCI for each group of experiments. We also averaged the scores of all Likert items to produce an overall score—arguably a score of creativity. These results, with standard error, can be seen in Figures 6.10-6.15.

In order to discover the effect of bias on the results in the second part of the survey, we calculated the Pearson correlation coefficient, $r$, between bias and the average Likert item scores for both informed and blind experiments. A positive correlation between bias and a particular item for a particular artist (either human or DARCI) would indicate that a bias towards humans (or against DARCI) is correlated with an increase in the item score for the particular artist. Tables 6.1 and 6.2 show these correlation values and accompanying $p$-values for the informed and blind experiments respectively.

In the informed experiment, bias correlated with an increase in scores towards human artists for the difficult ($r = 0.161$) and overall ($r = 0.144$) items. This correlation was statistically significant to $p < 0.05$. While not all of the individual correlation values were statistically significant, bias trended towards positive correlation with human scores and negative correlation with DARCI scores (see Table 6.1). These results indicate that volunteers with a predisposed bias towards humans, or a bias against the DARCI project, increased
Table 6.1: The Pearson correlation coefficient, $r$, and associated $p$-value, between volunteer bias and item scores for the informed experiment. A positive correlation indicates that a bias towards humans is correlated with an increase in item score.

<table>
<thead>
<tr>
<th>Artist</th>
<th>Item</th>
<th>$r$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>like</td>
<td>0.067</td>
<td>0.364</td>
</tr>
<tr>
<td>human</td>
<td>adjective</td>
<td>-0.034</td>
<td>0.639</td>
</tr>
<tr>
<td>human</td>
<td>surprising</td>
<td>0.136</td>
<td>0.064</td>
</tr>
<tr>
<td>human</td>
<td>difficult</td>
<td>0.161</td>
<td>0.028</td>
</tr>
<tr>
<td>human</td>
<td>use</td>
<td>0.096</td>
<td>0.191</td>
</tr>
<tr>
<td>human</td>
<td>overall</td>
<td>0.144</td>
<td>0.048</td>
</tr>
<tr>
<td>DARCI</td>
<td>like</td>
<td>-0.046</td>
<td>0.532</td>
</tr>
<tr>
<td>DARCI</td>
<td>adjective</td>
<td>-0.111</td>
<td>0.127</td>
</tr>
<tr>
<td>DARCI</td>
<td>surprising</td>
<td>0.006</td>
<td>0.930</td>
</tr>
<tr>
<td>DARCI</td>
<td>difficult</td>
<td>-0.047</td>
<td>0.517</td>
</tr>
<tr>
<td>DARCI</td>
<td>use</td>
<td>0.006</td>
<td>0.932</td>
</tr>
<tr>
<td>DARCI</td>
<td>overall</td>
<td>-0.059</td>
<td>0.424</td>
</tr>
</tbody>
</table>

some of the scores given to humans and *may* have decreased some of the scores attributed to DARCI.

While one might expect no correlation between bias and scores in the blind experiment, there was a clear trend towards negative correlation across *all* items, both for humans and DARCI (see Table 6.2). None of these correlation values were statistically significant, but the fact that almost all of the correlations were negative suggests that there may indeed be an overall negative correlation. This would imply that those with a bias towards humans, or against the DARCI project, tend to give all images a lower score when they don’t know who produced them. We suspect that this may be because these volunteers are considering the possibility that an image is produced by a computer program. It would be interesting to investigate this phenomenon in future studies.

Another item of note is the fact that there was a nearly statistically significant negative correlation between bias and correct answers to the comprehension question ($r = -0.181$, $p = 0.093$). This suggests that those biased towards humans, or against the DARCI project, tended to more often fail the comprehension question implying that they didn’t care about understanding how DARCI works as much. Again, this result was not statistically significant, so further study will be necessary to verify this observation.
<table>
<thead>
<tr>
<th>Artist</th>
<th>Item</th>
<th>r</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>like</td>
<td>-0.087</td>
<td>0.399</td>
</tr>
<tr>
<td>human</td>
<td>adjective</td>
<td>-0.089</td>
<td>0.389</td>
</tr>
<tr>
<td>human</td>
<td>surprising</td>
<td>-0.043</td>
<td>0.677</td>
</tr>
<tr>
<td>human</td>
<td>difficult</td>
<td>0.019</td>
<td>0.851</td>
</tr>
<tr>
<td>human</td>
<td>use</td>
<td>-0.154</td>
<td>0.133</td>
</tr>
<tr>
<td>human</td>
<td>overall</td>
<td>-0.127</td>
<td>0.216</td>
</tr>
<tr>
<td>DARCI</td>
<td>like</td>
<td>-0.047</td>
<td>0.651</td>
</tr>
<tr>
<td>DARCI</td>
<td>adjective</td>
<td>-0.076</td>
<td>0.462</td>
</tr>
<tr>
<td>DARCI</td>
<td>surprising</td>
<td>-0.045</td>
<td>0.661</td>
</tr>
<tr>
<td>DARCI</td>
<td>difficult</td>
<td>-0.098</td>
<td>0.342</td>
</tr>
<tr>
<td>DARCI</td>
<td>use</td>
<td>-0.073</td>
<td>0.480</td>
</tr>
<tr>
<td>DARCI</td>
<td>overall</td>
<td>-0.103</td>
<td>0.316</td>
</tr>
</tbody>
</table>

Table 6.2: The Pearson correlation coefficient, $r$, and associated $p$-value, between volunteer bias and item scores for the blind experiment. A positive correlation indicates that a bias towards humans is correlated with an increase in item score.

6.5.2 Evaluation of Creativity

Looking at Figures 6.10-6.15 we see that DARCI scored higher than human artists overall and with the surprising and difficult categories while humans scored higher than DARCI with the like and use categories. With the adjective category, both DARCI and human artists scored similarly. These trends persisted across all experiments despite the overall human bias of the volunteers. While many of the comparisons between DARCI and human artists were close to being statistically significant ($p < 0.05$), only those starred (*) were actually statistically significant. From a purely statistical perspective, DARCI exceeded human performance in the overall, surprising, and difficult categories while human artists did not exceed DARCI’s performance in any category.

While purely quantitative, these results suggest that within this constrained domain of digital visual art, DARCI is capable of producing renderings that are comparable to human renderings in terms of appeal, while being significantly more surprising and unusual. This is more than just a functional evaluation of DARCI’s artifacts, it’s also an evaluation of the creation process. The fact that DARCI scored higher than humans in the difficulty category suggests that volunteers felt that DARCI’s artifacts required some skill to create. Additionally, volunteers who passed the comprehension question appreciated DARCI’s artifacts as well as
Figure 6.10: The average scores of all Likert items across all artifacts produced by either humans or DARCI for each group of experiments (with standard error). (*) indicates statistical significance between human and DARCI results.

<table>
<thead>
<tr>
<th>Like</th>
<th>Cold</th>
<th>Eerie</th>
<th>Violent</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.873</td>
<td>2.260</td>
<td>1.870</td>
<td>1.377</td>
</tr>
</tbody>
</table>

Table 6.3: The four Likert scores for the source image.

those who did not pass the question—understanding how DARCI functioned did not diminish the way the artifacts were perceived.

Table 6.3 shows the average scores of the source image for its four Likert items across all experiments. While survey takers liked the source image, they did not think that it reflected any of the adjectives particularly well. Comparing Table 6.3 with Figure 6.12 we see that both humans and DARCI were able to greatly improve the adjective scores with their artifacts.

There was no statistically significant difference between the results of any of the four experimental groups (blind, informed, basic, and detailed), except between the basic and detailed groups in the adjective category (see Figure 6.12). Understanding how DARCI produced artifacts clearly influenced how volunteers perceived the meaning of the images produced by both DARCI and humans. While we can’t be certain of the reason, this change in perception towards both DARCI and humans may have been because it was indicated to
Figure 6.11: The average scores of the like item across all artifacts produced by either humans or DARCI for each group of experiments (with standard error). (*) indicates statistical significance between human and DARCI results.

Table 6.4: The top six ranked images for each Likert item for the informed experiment. Refer to Figures 6.7 and 6.8 to view images.

<table>
<thead>
<tr>
<th>Like</th>
<th>Adjective</th>
<th>Surprising</th>
<th>Difficult</th>
<th>Use</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human 1 “cold”</td>
<td>DARCI 4 “cold”</td>
<td>DARCI 2 “violent”</td>
<td>DARCI 1 “violent”</td>
<td>Human 1 “cold”</td>
<td>DARCI 4 “cold”</td>
</tr>
<tr>
<td>DARCI 3 “cold”</td>
<td>DARCI 1 “cold”</td>
<td>Human 3 “eerie”</td>
<td>Human 3 “eerie”</td>
<td>DARCI 1 “cold”</td>
<td>DARCI 1 “cold”</td>
</tr>
<tr>
<td>Human 4 “violent”</td>
<td>Human 2 “cold”</td>
<td>DARCI 2 “eerie”</td>
<td>DARCI 2 “eerie”</td>
<td>DARCI 1 “cold”</td>
<td>DARCI 2 “eerie”</td>
</tr>
<tr>
<td>DARCI 2 “eerie”</td>
<td>Human 2 “eerie”</td>
<td>DARCI 2 “violent”</td>
<td>Human 2 “violent”</td>
<td>Human 4 “eerie”</td>
<td>DARCI 3 “cold”</td>
</tr>
<tr>
<td>Human 4 “eerie”</td>
<td>DARCI 4 “cold”</td>
<td>DARCI 4 “cold”</td>
<td>DARCI 4 “violent”</td>
<td>Human 4 “cold”</td>
<td>DARCI 4 “eerie”</td>
</tr>
</tbody>
</table>

In other words, we suggest that volunteers may have been incorporating Jordanous’ ‘social interaction and communication’ component into their evaluation.

Tables 6.4 and 6.5 show the top six rated images for each category for informed and blind experiments. Refer to Figures 6.7 and 6.8 to view the actual images. Of note, DARCI’s artifacts have better representation amongst these highest rated images.
Figure 6.12: The average scores of the *adjective* item across all artifacts produced by either humans or DARCI for each group of experiments (with standard error). (*) indicates statistical significance between human and DARCI results.

Table 6.5: The top six ranked images for each Likert item for the *blind* experiment. Refer to Figures 6.7 and 6.8 to view images.

### 6.6 Conclusions

We have described a computational system, DARCI, that generates renderings of images so that they reflect a given adjective. The system is designed to produce “creative” artifacts through a “creative” process. We have also presented a human survey designed to evaluate DARCI’s artifacts and creation process while taking participant bias into consideration. The survey also uses human artists’ artifacts as a baseline for analyzing DARCI’s results. Such a survey could be generalized to many computational systems, though it would need to be tailored to the specific domain of the system in question.

By analyzing the survey results, and considering each Likert item as a component of creativity, we have shown that DARCI’s artifacts are statistically significantly at least as “creative” as those produced by humans. This performance over humans comes primarily
Figure 6.13: The average scores of the surprising item across all artifacts produced by either humans or DARCI for each group of experiments (with standard error). (*) indicates statistical significance between human and DARCI results.

from the surprising nature of DARCI’s images and the perceived difficulty in creating them. DARCI’s performance in the evaluation persisted even when survey takers (shown to be biased against DARCI) were aware of the process used to create the images.

While these results look remarkable on paper, we must note that creativity is still ill-defined and our survey questions are clearly a simplification of what it means to be creative (hence the quotes earlier). We must also acknowledge that the artifacts were very specific in nature and that the human artists’ creative process was heavily restricted in order to make the comparison to DARCI fair. In a more practical setting, humans would have far fewer restrictions and would undoubtedly produce more interesting images. Finally, we must acknowledge that the four sets of DARCI’s artifacts used in the survey were selected from seven sets by a human—though more than half of DARCI’s artifacts were included.

Despite these limitations, the results clearly indicate a system capable of performing on par with humans within the restricted domain. These results will also act as a valuable baseline for testing future improvements to the system.
Figure 6.14: The average scores of the difficult item across all artifacts produced by either humans or DARCI for each group of experiments (with standard error). (*) indicates statistical significance between human and DARCI results.

Figure 6.15: The average scores of the use item across all artifacts produced by either humans or DARCI for each group of experiments (with standard error). (*) indicates statistical significance between human and DARCI results.
Chapter 7

Improving Affective Image Annotation with Features that Summarize Local Interest Points¹

Abstract

With the growing prevalence of media sharing websites, the detection of abstract concepts in images, particularly emotions, is an increasingly important capability of image retrieval systems. We present a set of image features that combines established features from state-of-the-art approaches with novel features that summarize local interest-point data. Using three established datasets spanning eight affective labels, we show that our approach improves upon the state-of-the-art in most cases. While much research is being done in this area, most is limited to oversimplified sets of affective labels. To overcome the limited range of affective labels, we introduce a new dataset for training and evaluating image annotation systems across a large range of WordNet adjectives, including many affective adjectives. Using this dataset, we demonstrate results across 110 adjective labels comparable to results across the aforementioned eight affective labels. While even our improved results are not ideal for most practical applications, we present arguments to reevaluate assumptions made about datasets that suggest a more optimistic future for affective semantic image retrieval.

¹David Norton, Derrall Heath, and Dan Ventura. Improving affective image annotation with features that summarize local interest points. In submission to the IEEE Transactions on Affective Computing, 2014.
7.1 Introduction

As information becomes increasingly abundant, the need for effective information retrieval techniques is become more and more important. With the prevalence of online media sharing sites such as Facebook, Google Images, and Flickr, images have become a particularly important medium for information retrieval. Such venues have historically offered effective image retrieval primarily through the matching of search terms to the contextual text of images: tags, titles, captions, HTML code, proximal text, etc. While valuable, much of the information expressed with an image is not contained in the proximate text. The ability to automatically annotate images with descriptors based entirely on image content would improve the range and reliability of image retrieval algorithms and is the subject of Content Based Image Retrieval (CBIR) research. While much research in the domain of CBIR is focused on identifying the subject of images, the sub domain of Emotional Semantic Image Retrieval (ESIR) is interested in identifying the emotional impact of images.

ESIR research is not only valuable in the retrieval of images, it is also valuable in the domain of computer generated art—art that is fully or semi-autonomously created with a computer program. Computer generated art is useful to the visual arts as both a powerful tool for creation and a source of inspiration. More autonomous varieties of computer generated art also have potential in advertising, interactive narrative, and game design by generating images for a particular context or emotion on the fly. Research surrounding computer generated art falls within the purview of computational creativity, the growing subfield of artificial intelligence interested in the automation of tasks that could be said to require creativity. Examples of “creative” systems designed to produce visual art include Harold Cohen’s AARON [61], Simon Colton’s The Painting Fool [13], DiPaola and Gabora’s Evolving Darwin’s Gaze [27], Machado et al.’s NEvAr [57], and our own system DARCI [70].

DARCI (Digital ARtist Communicating Intention) is designed to produce original digital images that communicate meaning to a viewer through the use of image filters. Like many other digital “artists”, DARCI uses an evolutionary algorithm to generate images. To
be fully autonomous, this essentially requires that the system evaluate its own creations and iteratively modify them until a satisfactory threshold is reached. The central component of this evaluation is a measure of how adequately an image reflects intended meaning. In our current research, we have restricted this meaning to adjectives with an emphasis on affective adjectives. This requires the extensive use of ESIR approaches to annotate images with appropriate labels. Unfortunately, ESIR research has had minimal success due to the inherent challenges of feature selection in image annotation, subjectivity regarding the emotional impact of images, and limited availability of appropriate datasets. Furthermore, most existing research in ESIR is restricted to a simplified range of affective categories, a limitation we want to avoid with DARCI.

In this paper, we present a set of image features that combines established features from state-of-the-art ESIR approaches with new features, particularly those that summarize local interest-point data. We also present a new dataset for training and evaluating image annotation systems across a large range of adjectives, including affective adjectives. All of our classification results are obtained using binary artificial neural networks trained with standard back propagation.

We begin by reviewing several state-of-the-art approaches to ESIR and other related forms of image annotation. We then compare and contrast three established datasets with our new dataset, illustrating the need for datasets like ours. Following this, we describe 197 image features and our feature selection approach. Finally, we evaluate our features and model, and compare our results to those of several other state-of-the-art methods. We show that our model and features improve upon the state-of-the-art in most cases.

7.2 Background

Image annotation research is typically focused on the detection of objects in images: faces, cars, animals, etc. Yet, there is a need for the annotation of more abstract concepts in images,
such as emotional content, image genre, and aesthetic quality. While less common, several attempts have been made at more abstract annotation in the past ten years.

Automatically assessing the aesthetic quality of images is of interest to artists and photographers looking to receive quick, unbiased feedback about their work. Two research groups have led the way in this type of annotation. Datta et al. use global image features to assess the aesthetic quality of images based on training from peer-rated online photo sharing websites [24]. Li et al. combine global features with segmentation features to assess the aesthetic quality of paintings [53].

Detecting image genre is another form of abstract image annotation. Zujovic et al. use color and gray-level features along with various machine learning algorithms to classify images of paintings into one of five different genres: Abstract Expressionism, Cubism, Impressionism, Pop Art, and Realism [97].

Sentiment analysis is a popular topic in data mining and web page analysis. Such analysis can be enhanced by the assessment of sentiment in images based on image content. Using high-dimensional local feature histograms, Siersdorfer et al. and Borth et al. assess the sentiment of images on Flickr and Twitter respectively [9, 82].

Finally, several researchers have examined various methods and features for detecting the emotional impact of images. Wang et al. build classifiers that learn to identify 12 dichotomous pairs of emotional words using three specially designed sets of global features [92]. Dellagiacoma et al. use psychological studies to select ideal color and texture features for classifying images into five emotional categories: surprise, fear, disgust, happiness, and sadness [26]. They eliminate content bias by only classifying photographs containing natural bodies of water.

None of the datasets used for training and testing in the aforementioned papers have been made available, making direct comparison difficult. In contrast, Yanulevskaya et al. use SVMs to classify eight emotions in a dataset of images that is well established in the field of psychology [94]. Machajdik and Hanbury make available two additional datasets and perform
a thorough analysis of image features used throughout ESIR research [58]. They use a naïve Bayes classifier to categorize the same eight emotions across these three datasets.

The described research has provided a valuable look at myriad features for annotating images with a variety of abstract concepts. However, none of these studies has examined beyond a selected handful of concepts; nor have any of them accounted for the realistic scenario of labeling an image with multiple concepts. In this paper, we build on this research by employing many of the features described while adding new ones. In addition, we annotate images with a wider range of labels and account for multiple labels per image.

7.3 Datasets

To evaluate and critique our approach in this paper and to facilitate comparison to other approaches, we employ four datasets: IAPS, Art Photo, Abstract, and DARCI. The first three datasets come from previous research in the Emotional Semantic Image Retrieval (ESIR) community [58, 94]. The fourth is a dataset that we have developed and present here as a complementary addition.

7.3.1 IAPS

The IAPS (International Affective Picture System) dataset is a popular resource used in research dealing with emotions [52]. It contains 1196 images with subject ratings across three dimensions: affective valence, arousal, and dominance. Affective valence is the positive or negative affect an image has on a viewer, arousal is the degree of excitement that an image incites in a viewer, and dominance is how in control the image makes the viewer feel. Mikels et al. categorized a subset of the IAPS dataset into 8 affective categories divided into those with positive and those with negative valence [62]. The positive valence categories are amusement, awe, contentment, and excitement. The negative categories are anger, disgust, fear, and sadness. To determine the 4 categories in each valence, Mikels performed a pilot study allowing free response to test images. The resulting affective categories were the top
results from the pilot study. Mikels et al. then performed two studies (one for positive emotions and one for negative emotions) to determine how participants attributed (on a 7 point scale) the 8 emotions to 390 of the IAPS images. The resulting dataset is the *IAPS* dataset we use here.

The primary strengths of this dataset are that it is based on rigorous psychological experimentation and contains real valued scores for each image-emotion pair. Additionally, Mikels et al. use confidence intervals to threshold the real valued scores and explicitly determine which labels apply to which images and which do not. These label assignments allow for images to be labeled with multiple emotions and eliminate the need to guess at negativity.

One weakness of this dataset is that the emotional content of the images is primarily determined by their subject as opposed to their presentation and other global factors. The IAPS images typically contain one or two central objects of interest on which most of the emotional reaction is dependent. This makes for a particularly challenging problem within the scope of global image features. Another weakness is that the *IAPS* dataset categorized by Mikels et al. contains only 8 images labeled with anger making that category almost useless.

The original IAPS dataset is available online by request\(^2\). The category data determined by Mikels et al. is also available online as supplementary material for their paper\(^3\).

### 7.3.2 Art Photo

The *Art Photo* dataset consists of 806 photographs Machajdik and Hanbury collected from an art sharing site\(^4\) by searching for images that were labeled by their author with one of the above 8 affective categories [58]. Thus, this dataset reflects the intent of the author in the images. Unfortunately, this dataset only contains a single affective label for each image. This

\(^2\)http://csea.phhp.ufl.edu/media.html  
\(^3\)http://www.psychonomic.org/behavior-research-methods  
\(^4\)www.deviantart.com
makes it unreliable to assume negative examples for binary classification. This dataset is made available by Machajdik and Hanbury online\(^5\).

### 7.3.3 Abstract

The Abstract dataset contains 280 images of abstract paintings. Machajdik and Hanbury created this dataset by polling roughly 230 volunteers with a web-survey \([58]\). Volunteers were asked to label 20 random images with one of the above 8 emotions. The result was about 14 labels per image. Since the images are abstract, a fair assessment of global image features can be made. Unfortunately, even though multiple users labeled each image, each user was restricted to labeling images with a single emotion—an often unrealistic restriction as many of these abstract images are emotionally ambiguous. This dataset is also made available by Machajdik and Hanbury online\(^4\).

### 7.3.4 DARCI

Realistically, the annotation of images with emotions is a multi-label problem where implicit negativity cannot be assumed. In other words, images can contain multiple emotions; and being labeled with one emotion does not automatically imply that the image does not reflect any other emotions. Additionally, there is a wide spectrum of human emotions that cannot be fully encompassed by 8 concepts. Shoehorning images into a small number of emotions will not necessarily yield the impact an image has on a particular viewer. The complexity of the emotion annotation problem reveals the draw backs of the three datasets just described.

For the past four years we have been developing a dataset of labeled images to address these problems and train our creative system, DARCI. This dataset is continuously growing as volunteers add datapoints and new images via a website\(^6\). Images range from drawings, to paintings, to photographs, to abstract digital renderings—there are no restrictions on image

\(^5\)http://www.imageemotion.org/

\(^6\)http://darci.cs.byu.edu/
Figure 7.1: Screen capture illustrating one interface for labeling images for the DARCI dataset. Volunteers are prompted to label the image with any adjectives.

 Volunteers are allowed to label images with any number of adjectives existing in the WordNet\textsuperscript{7} database—over 18,000.

 In order to eliminate ambiguity, when volunteers type in labels for images, they are asked to indicate the particular sense of the word they intend; thus, labels are technically WordNet adjective synsets (unique qualifiers for specific definitions). Figure 7.1 illustrates how labeling images is currently handled. Since volunteers are free to use almost any adjective, most affective concepts can be expressed when describing each image. Furthermore, through a variety of labeling exercises, volunteers can specify labels that explicitly do not describe the images. In this way, explicit negative labels are obtained—a valuable asset for machine learning algorithms. Figure 7.2 illustrates one method of obtaining negative datapoints.

 Of course, with the scope of this dataset comes its own set of weaknesses. Most prominent, the data is not collected in a controlled environment. In fact, the interface for obtaining the labels has even changed over time \textsuperscript{66}. While numerous individuals have no

\footnote{http://wordnet.princeton.edu/wordnet/}
Figure 7.2: Screen capture illustrating one interface for obtaining negative datapoints for DARCI dataset. Volunteers are prompted to remove adjectives that they think are wrong.

doubt provided labels, a large number of the datapoints were collected from people closely associated with the DARCI project itself. It is also possible that attempts have been made to sabotage the data. However, the website has not reached a level of popularity so as to invite such behavior, nor have we detected it in the data.

In addition to this admittedly significant drawback, is the problem of data scarcity. With an ever growing bank of images, and over 18,000 potential labels, the dataset is arguably sparse. That being said, there are 71 adjectives, many of which are affective, with over 100 unique image associations. This subset of datapoints alone represents a more diverse and densely populated dataset then the three described previously. More statistics about the DARCI dataset can be seen in Table 7.1. The dataset itself can be found online\textsuperscript{5}.

The more controlled environments of the first three datasets, and the fact that they cover the same 8 emotions, will ensure that these datasets continue to be valuable to the ESIR community. However, due to its breadth and detail, we suggest that the DARCI dataset can be another valuable dataset for ESIR research. All four datasets are used in this research.
### Table 7.1: Statistics about the DARCI dataset

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total datapoints</td>
<td>33129</td>
</tr>
<tr>
<td>Total positive datapoints</td>
<td>17004</td>
</tr>
<tr>
<td>Total negative datapoints</td>
<td>16125</td>
</tr>
<tr>
<td>Total adjectives</td>
<td>2463</td>
</tr>
<tr>
<td>Total images</td>
<td>2562</td>
</tr>
<tr>
<td>Average number of unique adjectives per image</td>
<td>11.78</td>
</tr>
<tr>
<td>Number of adjectives used as labels for least 30 unique images both positively and negatively</td>
<td>110</td>
</tr>
</tbody>
</table>

#### 7.4 Image Features

In previous work we have relied on a set of 102 global image features, summarized in Table 7.2 [66]. Most of these features come from the DISCOVIR (Distributed Content-based Visual Information Retrieval) system package available for Java\(^8\). DISCOVIR is designed as a shareable resource for content-based image retrieval on the internet [51]. While many of the features available in the package are useful for this purpose, many are outdated or inefficient for our needs.

In this section we explore the usefulness of the DISCOVIR features we have used in the past, in addition to many new features from a variety of sources including some original to this work. We will begin by describing features that pertain to color, followed by features that pertain to textures and patterns, followed by new global features derived from local image features. At the end of this section, we will describe the feature selection process we used to determine the best set of these features for universal adjective annotation. As we describe each feature, we will provide a label for that feature that will be used throughout the rest of the paper for easy identification.

#### 7.4.1 Color Features

Since we are primarily interested in the emotional content of images, features that have been shown to correlate with the emotional content of images would be most relevant to our work.\(^8\)

\(^8\)Website: [http://appsrv.cse.cuhk.edu.hk/~miplab/discovir/](http://appsrv.cse.cuhk.edu.hk/~miplab/discovir/)
Table 7.2: Summary of features from prior work [66]

<table>
<thead>
<tr>
<th>Color &amp; Light</th>
<th>Texture</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Average red, green, and blue</td>
<td>1. Co-occurrence matrix (x4)</td>
</tr>
<tr>
<td>2. Average hue, saturation, and intensity</td>
<td>1. Maximum probability</td>
</tr>
<tr>
<td>3. Saturation and intensity contrast</td>
<td>2. First order element difference moment</td>
</tr>
<tr>
<td>4. Unique hue count (from 20 quantized hues)</td>
<td>3. First order inverse element difference moment</td>
</tr>
<tr>
<td>5. Hue contrast</td>
<td>4. Entropy</td>
</tr>
<tr>
<td>6. Dominant hue</td>
<td>5. Uniformity</td>
</tr>
<tr>
<td>7. Dominant hue image percent</td>
<td>2. Edge frequency (25x vector)</td>
</tr>
</tbody>
</table>

Based on our own findings [43] and research by Wang and Li [53, 92], many of these features focus on color. As such, we explore a variety of color features, 56 to be exact.

**Standard Color Space Statistics**

We examine several statistics derived from a variety of color spaces. We look at the average red, green, and blue pixel values from the RGB color space ($meanR$, $meanG$, $meanB$); the average hue, saturation, and intensity pixel values from the HSI color space ($meanH$, $meanS$, $meanI$); and the average lightness, chroma, and hue pixel values derived from the CIELAB color space ($meanL$, $meanC$, $meanh_{ab}$). The average RGB and HSI values are taken from the DISCOVIR package. We add features from the CIELAB color space because this space has been shown to most accurately map to the way our eyes see lightness and color [29].

While distinct features, the hue value in both the HSI and CIELAB color space are angular in nature and thus require circular statistics to accurately calculate their averages. For better or for worse, the DISCOVIR package uses the arithmetic mean rather than the
circular mean to calculate average hue values from the HSI color space. However, for hue
from the CIELAB color space, we use the circular mean calculated by:

\[ h_\mu = \text{atan2} \left( \frac{1}{n} \sum_{i=1}^{n} \sin h_i, \frac{1}{n} \sum_{i=1}^{n} \cos h_i \right) \] (7.1)

where \( h \) is the hue angle, \( n \) is the total number of pixel values, and the function \( \text{atan2}(y, x) \)
computes the angle of the coordinates \( x \) and \( y \) to be in the range \( (-\pi, \pi] \). The circular mean
essentially converts the angular values into Cartesian coordinates, calculates the arithmetic
mean, then returns the angle of the new coordinates.

In addition to averages, we calculate the standard deviation of lightness and chroma
along with the hue spread from the CIELAB space (\( \text{stdL}, \text{stdC}, \text{spread}_{ab} \)). Hue spread is
another example of a circular statistic which tells us the range of hue pixel values present in
the image. It is calculated by:

\[ h_\sigma = 1 - \sqrt{\left( \frac{1}{n} \sum_{i=1}^{n} \sin h_i \right)^2 + \left( \frac{1}{n} \sum_{i=1}^{n} \cos h_i \right)^2} \] (7.2)

Finally, using DISCOVIR we calculate the contrast of saturation and intensity from
the HSI color space (\( \text{contrastS}, \text{contrastI} \)).

**Emotional Color Space Statistics**

Ou [74] performed psychological experiments that essentially derive an emotional color space
from the CIELAB color space. This emotional color space consists of models of three
emotional factors identified using factor analysis on ten color-emotional scales: likedislike,
fresh-stale, clean-dirty, modern-classical, hard-soft, tense-relaxed, masculine-feminine, heavy-
light, warm-cool, and active-passive. The three models are color activity, weight, and heat.

Color activity essentially describes the distance of a pixel color from a medium grey,
and includes features from the active-passive, fresh-stale, clean-dirty, and modern-classical
scales. It is calculated by:

$$\textit{activity} = -2.1 + 0.06 \left[ (L^* - 50)^2 + (a^* - 3)^2 + \left( \frac{b^* - 17}{1.4} \right)^2 \right]^{1/2}$$  \hspace{1cm} (7.3)$$

where $L^*$ is lightness from the CIELAB color space and $a^*$ and $b^*$ are coordinates from the CIELAB color space.

Color weight describes the richness of pixel colors and includes features from the hard-soft, masculine-feminine, and heavy-light scales. It is calculated by:

$$\textit{weight} = -1.8 + 0.04(100 - L^*) + 0.45\cos(h - 100^\circ)$$  \hspace{1cm} (7.4)$$

where $h$ is hue from the CIELAB color space.

Color heat describes the color temperature of pixels and includes only features from the warm-cool scale. It is calculated by:

$$\textit{heat} = -0.5 + 0.02(C^*)^{1.07}\cos(h - 50^\circ)$$  \hspace{1cm} (7.5)$$

where $C^*$ is chroma from the CIELAB color space.

For our features, we calculate the average heat, weight, and activity pixel values in an image ($\textit{meanHeat}$, $\textit{meanWeight}$, $\textit{meanActivity}$) along with their standard deviations ($\textit{stdHeat}$, $\textit{stdWeight}$, $\textit{stdActivity}$).

**Rule of Thirds**

Inspired by the research of Datta, Li, and Machajdk, we extract some features from the central third region of the image to contrast with those from the entire image [24, 53, 58]. These features are added to provide some information as to how the rule of thirds is being followed in an image. The rule of thirds states that if we were to divide an image into 9 equal rectangles, the object of focus should exist somewhere on the edge of the middle rectangle. Since the object will often extend outside that third box, we analyze a region slightly larger
Figure 7.3: A sample image illustrating aspects of some of the features explored in this paper. (a) The original image. (b) The blue rectangle represents the center third of the image; the white one represents the part of the image used to calculate rule of thirds based features. (c) The CIELAB color space of the image discretized into 12 buckets (plus grey) according to the color histogram features. (d) The SURF interest points detected in the image with proper orientation and relative scale. The interest points have been shrunk in this image to improve figure clarity.

than the middle rectangle. Specifically, we divide the image into 16 equal rectangles and collect features from the middle 4 rectangles as shown in Figure 7.3(b).

The color features we collect from this middle section are the average lightness, chroma, and hue pixel values from the CIELAB color space ($focusMeanL$, $focusMeanC$, $focusMeanh_{ab}$), as well as the hue spread of the pixels within this area of focus ($focusSpreadh_{ab}$).
Colorfulness

To describe the use of color throughout the image we include four features that we call colorfulness features. These features, while not part of DISCOVIR, were part of the original 102 images features used in previous work [66].

These four features are a variation of the three colorfulness features described in [53] (features f1, f2, and f3 in that paper). For all of the colorfulness features, the first step is to create a 20-bin hue histogram. Hue is calculated from the HSI color space, and only those pixels with a saturation greater than 50 and an intensity between 38 and 242 are counted (here we have normalized all HSI values to range between 0 and 255). Pixels that fall outside of these specifications are indistinguishable from shades of grey. The range of hue values is evenly divided between the 20 histogram buckets and each qualifying pixel is added to the appropriate bucket.

The first feature is the index of the bin with the highest count (dominantHue). The second feature is the percentage of the image’s pixels that belong to the bin with the highest count (dominantHuePercent). The third feature is the number of bins that contain counts over some threshold (hueCount). In our paper, this threshold is $0.01 \cdot Q$, where $Q$ is the size of the largest bin in the histogram. The fourth and final colorfulness feature is the average contrast of the hue buckets (hueContrast) calculated as follows:

$$\sum_{i=1}^{20} \frac{\max_{h}(i) - \min_{h}(i)}{20}$$

(7.6)

where $\max_{h}(i)$ is the maximum hue value assigned to bin $i$ and $\min_{h}(i)$ is the minimum hue value designated to bin $i$.

Wang Emotional Histograms

In order to classify images with twelve emotional word-pairs, Wang et al. performed factor analysis to group the twelve word-pairs into three emotional categories [92]. Category one
includes the word pairs: exhilarated-depressive, warm-cool, happy-sad, light-heavy, hard-soft, brilliant-gloomy, and lively-tedious. Category two includes the word pairs: magnificent-modest, vibrant-desolate, and showy-elegant. Finally, category three includes the word pairs: clear-fuzzy and fanciful-realistic. For each of these categories, Wang et al. designed an aggregate set of features inspired by psychological research to predict the appropriate category of a given image. In their research, they showed a strong correlation between these features and how humans classified images.

We include a selection of Wang et al.’s features with our color features. Specifically, we use the 10-bin lightness/warmth histogram they designed for their first emotional category and the 6-bin saturation/warmth histogram they designed for their second emotional category. The lightness/warmth histogram identifies the fuzzy membership of an individual pixel in one of five lightness bins each within either a cool or warm bin making ten features (warmVeryDark, warmDark, warmMediumLight, warmLight, warmVeryLight, coolVeryDark, coolDark, coolMediumLight, coolLight, coolVeryLight). The saturation/warmth histogram identifies the fuzzy membership of an individual pixel in one of three saturation bins each within either a cool or warm bin making six features (warmDesaturated, warmMiddleSaturation, warmSaturated, coolDesaturated, coolMiddleSaturation, coolSaturated). The details of these histogram membership functions can be found in Wang et al.’s paper [92].

**Color Histogram**

The final set of color features we use in this paper is a 12-bin hue histogram. While essentially a measure of hue counts, this histogram differs from the 20-bin colorfulness histogram described previously in several ways. Firstly, this histogram is created with the CIELAB color space. Secondly, each bin in the histogram represents a unique color feature. Finally, the histogram uses a fuzzy membership function not unlike those of the Wang emotional histograms.

The first ten bins in the color histogram are for chromatic colors (magentaRed, redOrange, yellowOrange, yellowGreen, green, greenBlue, brightBlue, blue, purple, magenta).
Bin 11 and 12 are for white and black respectively \( (white, black) \). While the bin that a pixel is assigned is primarily determined by hue \( (h) \), chroma \( (C) \) and lightness \( (L) \) play a role in determining whether an image is achromatic: white, black, or grey (absent from the histogram). Since the histogram is fuzzy, edge cases are distributed between multiple bins. Figure 7.3(c) illustrates how colors are discretized in an example image. The color histogram is formally created as follows.

First we determine the chromatic factor, \( \chi \), or what fraction of the pixel will be assigned to one of the ten chromatic bins. At the same time, we determine to which achromatic color the remaining fraction, \( 1 - \chi \), will be assigned. If \( L < 6 \) then the achromatic color is black and \( \chi \) is determined by Equation 7.7. Otherwise, if \( C < 20 \) then the achromatic color is either white or grey and \( \chi \) is determined by Equation 7.8. In this case, if \( L > 90 \) then the achromatic color is white, otherwise it’s grey and not assigned a bin. In all other cases \( \chi = 1 \) and there is no achromatic color assigned to the pixel.

\[
\chi = \begin{cases} 
\max(\frac{L}{3} - 1, 0) \cdot \max(\frac{C}{10} - 1, 0) & \text{if } C < 20 \\
\max(\frac{L}{3} - 1, 0) & \text{if } C \geq 20
\end{cases}
\]  

\( (7.7) \)

\[
\chi = \begin{cases} 
\max(\frac{C}{10} - 1, 0) \cdot \max(19 - \frac{L}{5}, 0)) & \text{if } L > 90 \\
\max(\frac{C}{10} - 1, 0) & \text{if } L \leq 90
\end{cases}
\]  

\( (7.8) \)

In other words, if the lightness is low enough, then the pixel’s color is considered partially black—the fraction of blackness is greater the darker and more desaturated the pixel is. Otherwise, if a pixel is desaturated enough, then a light pixel will be considered white while a darker pixel will be considered grey. The thresholds for these delineations were determined through observation of the perceived colors of pixel values.

Once \( \chi \) is determined, we calculate the hue factor, \( \eta^{bin} \), for each chromatic bin using the function in Figure 7.4 based on \( h \). The membership function for each chromatic bin is then
Figure 7.4: A function used to determine $\eta^{\text{bin}}$ (vertical axis) for a given hue, $h$ (horizontal axis). Each bin is labeled with its color name, with an approximation of the represented color below. As a clarifying example highlighted in the Figure: for a pixel with $h = 288$, $\eta^{\text{blue}} = 0.28$, $\eta^{\text{purple}} = 0.72$, and all other bin values are 0.

calculated by: $\chi \cdot \eta^{\text{bin}}$. In review, the membership function for the appropriate achromatic bin is: $1 - \chi$; if the achromatic color is grey, this bin is functionally outside of the histogram.

7.4.2 Global Texture and Shape Features

The DISCOVIR features extraction package contains 90 features for analyzing the texture and shape patterns found in images that we examine here. We will briefly summarize these, but details and source code can be found at the DISCOVIR website\textsuperscript{7}.

Co-Occurrence

A co-occurrence matrix of gray scale pixel values is calculated for an image at four different shifts (1,4,9,16). Each shift determines the distance and direction, in number of pixels along both the $x$- and $y$-axis, of the two pixels being compared to determine co-occurrence. Using these matrices, five features are extracted for each shift for a total of 20 co-occurrence features: maximum probability, first order element difference moment, first order inverse element difference moment, entropy, and uniformity ($c\text{MaxProbability1}$, $c\text{Moment1}$, $c\text{InverseMoment1}$, $c\text{Entropy1}$, $c\text{Uniformity1}$, $c\text{MaxProbability4}$, $c\text{Moment4}$, ... $c\text{Entropy16}$, $c\text{Uniformity16}$).
**Edge Frequency**

We use DISCOVIR to calculate edge frequencies at 25 different scales. Each scale defines the distance in pixels between a given pixel and its four orthogonal neighbor pixels. Edge frequency is calculated as the mean gradient of every pixel in an image compared to its four orthogonal neighbors. The 25 scales range in 2 pixel increments from 1 pixel to 49 pixels (edgeFreq1, edgeFreq3, edgeFreq5, ... edgeFreq49).

**Primitive Length**

DISCOVIR primitives are defined as continuous regions of pixels with the same gray scale value running in the same direction. The number of primitives of varying length from 1 to $MAX(\text{imagewidth}, \text{imageheight})$ for each of 256 gray scale values is calculated. Five features are extracted from these statistics: a quantity emphasizing short primitives, a quantity emphasizing long primitives, the amount of gray-scale uniformity, the amount of primitive length uniformity, and the percentage of the image occupied by primitives (shortPrimitives, longPrimitives, grayUniformity, primitiveUniformity, primitivePercent).

**Image Moments**

The DISCOVIR feature extraction package calculates several image moments as features to describe the shape of the image. We use the package to calculate the raw moment $M_{11}$ called the geometric moment ($geometricM$). We also use it to calculate the image eccentricity using raw moments ($eccentricity$). This indicates how elongated the image is. In addition, we use the DISCOVIR package to calculate five Hu invariant moments: $I_1, I_2, I_4, I_5, I_7$ (invariantM1, invariantM2, invariantM4, invariantM5, invariantM7). Finally, we use DISCOVIR to calculate the Legendre, Zernike, and Pseudo-Zernike moments ($LegendreM$, $ZernikeM$, $PseudoZernikeM$).
Edge Direction Histogram

We use DISCOVIR to build a 30 bin edge direction histogram. To build the histogram, the DISCOVIR feature extraction package determines the $x$- and $y$-gradient of each pixel in a given image blurred with a Gaussian kernel. It then uses these gradients to determine the gradient orientation at each pixel in the image and builds a uniform 30 bin histogram of these orientations ($\text{edge12}^\circ$, $\text{edge24}^\circ$, $\text{edge36}^\circ$, $\text{edge48}^\circ$, ... $\text{edge360}^\circ$).

7.4.3 Summary of Local Interest Point Features

While we are most interested in the global description of images with adjectives, we cannot escape the fact that the objects contained in an image impact the emotional content of the image. Unfortunately, universal object detection is an unsolved problem and thus clearly outside the scope of this paper. That being said, some of the most useful local image features in existing object detection algorithms may provide some added benefit to adjective classification.

In the area of object detection, the concept of visual words has seen success as a feature representation for images. Visual words are quantized region descriptors that act as the atomic unit of an image similar to how textual words act as the atomic unit of a document. In order to determine the visual words contained in an image, points of interest must first be identified followed by the salient region around the point. Typically, these interest points and salient regions are those that are the most surprising, or said another way, the least predictable. The Harris corner detector is one of the earliest successful interest point detectors and identifies the points in an image with the largest gradient surrounding the point [39]. This detector has been iterated upon numerous times with the intent of increasing invariance to various affine transformations. Along these lines, Lowe identified a scale invariant interest point detector using a difference of Gaussians function which, combined with a region descriptor, became SIFT (Salient Invariant Feature Transform) [55], the most commonly used descriptor in current object detection models.
In SIFT, once interest points are localized, each point is assigned an orientation based on surrounding gradients. Then a descriptor consisting of a gradient orientation histogram for the surrounding region is calculated. More recently, by focusing on essential properties, Bay et al. improved upon SIFT and other derived approaches, developing a faster and more robust approach called SURF (Speeded-Up Robust Features) [3].

Once interest point descriptors have been obtained, these features can be used to identify objects within and across images by essentially matching similar feature occurrences. In visual words models, a dictionary of visual words is created by extracting descriptors from a library of typical images. These descriptors, usually numbering in the hundreds of thousands, are then grouped into a number of clusters (usually thousands) using an unsupervised clustering algorithm such as $k$-means. Each of these clusters is a visual word, and all of the clusters constitute the dictionary for the application. When analyzing training and test images, every descriptor from each image is assigned to the closest visual word in the dictionary.

Akin to bag-of-words models in natural language processing, the frequency of visual words contained in an image can be recorded in a histogram—a bag-of-visual-words. Sivic and Zisserman were the first to treat visual words with NLP approaches and even coined the term visual word [84]. Concurrent with Sivic’s work, Csurka et al. used bags of visual words with naïve Bayes and SVM classifiers to a similar end—though they used the term bag of keypoints [23]. Later, Sivic et al. went on to explore using the bag of visual words model with two well-establish statistical natural language processing models, probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA) [85]. These two NLP approaches are successful at categorizing images into unspecified topics just as they are at categorizing documents [5, 44].

Unfortunately, in bag-of-visual-word models, each one of the potentially thousands of bins is essentially a unique feature. Even though the bag-of-visual-words summarizes the information contained in all of an image’s interest points, the feature space is still extremely
large. We can’t effectively combine such a large feature space with the global features we have described previously without diluting the less abundant global features. However, we can use bags-of-visual-words and other interest point statistics to selectively create a smaller subset of features that summarize the local features found within an image. In this section, we will describe 52 features—that we call SLIP (Summary of Local Interest Point) features—taken from an analysis of local image features. All of the features described here use the SURF interest point detector and SURF descriptors. Figure 7.3(d) illustrates SURF interest points for a sample image.

**Interest Point Statistics**

The first features we will examine summarize a variety of statistics about the SURF interest points identified in a given image. These features give a general sense of space and the distribution of objects in the image.

We look at the total number of interest points detected in the image (numInterestPoints). This tells us in general how busy the image is. We also identify the use of variable depth of field in images with a feature based on the low depth of field indicators developed by Datta et al. [24]. For this feature, we count the number of interest points detected in the middle section of the image as indicated in Figure 7.3(b), then take the ratio of this count over the total number of interest points detected in the image (depthOfField). This feature will tell us how focused the image is on central objects.

SURF interest points have an orientation and scale associated with them. We calculate the average orientation and scale of all interest points in the image (meanOrientation, meanScale). Since orientation is a polar value, we use circular statistics to calculate its mean (see Equation 7.1). We also calculate the orientation spread (see Equation 7.2) and standard deviation of scale (spreadOrientation, stdScale).

In addition to these mean and standard deviation values, we create small histograms of interest point orientations and scales. The orientation histogram consists of 8 bins
evenly divided across 360°, beginning with the bin that includes interest points oriented between 337.5° and 22.5° (orientation0°, orientation45°, orientation90°, orientation135°, orientation180°, orientation225°, orientation270°, orientation315°).

The scale histogram also consists of 8 bins. These bins are divided amongst the range of SURF interest point scale values as follows. Where \( \sigma \) is interest point scale, bin 1: \( \sigma < 1.93 \), bin 2: \( 1.93 \leq \sigma < 5.78 \), bin 3: \( 5.78 \leq \sigma < 9.63 \), bin 4: \( 9.63 \leq \sigma < 13.48 \), bin 5: \( 13.48 \leq \sigma < 17.33 \), bin 6: \( 17.33 \leq \sigma < 21.18 \), bin 2: \( 21.18 \leq \sigma < 25.03 \), bin 2: \( 25.03 \leq \sigma \) (scale1, scale2, scale3, scale4, scale5, scale6, scale7, scale8).

We calculate the average normalized \( x \) and \( y \) coordinates of each interest point along with their standard deviations (mean\( X \), mean\( Y \), std\( X \), std\( Y \)). In addition, we calculate the average Euclidean distance between all pairs of interest points as well as the standard deviation of these pairwise distances (mean\( \text{Distance} \), std\( \text{Distance} \)). Finally we calculate these same distance measures but normalized by the scale of the smallest interest point in a pair as Jamieson et al. do when calculating relationships between interest points [45] (mean\( \text{RelativeDistance} \), std\( \text{RelativeDistance} \)).

**Principal Components**

In order to assess the orientation of all of the interest points as a whole, we perform principal component analysis on the \( x \) and \( y \) dimensions of the interest points. We only use the eigenvector corresponding to the first principal component (principal\( X \), principal\( Y \)). We also calculate the strength of the principal component by dividing the eigenvalue of the first principal component by eigenvalue of the second component (principal\( \text{Strength} \)). This value of principal component strength tells us how concentrated the interest points are along the first principal component.
**SURF Descriptor Statistics**

SURF descriptors are a 64 feature vector. We extract the SURF descriptors for each interest point in the image and then calculate the mean and standard deviation of the difference between each pair of descriptors in the image. We do these calculations for three measures of difference: Euclidean distance, angular difference, and cosine similarity (\(mean_{Euclidean}, std_{Euclidean}, mean_{Angular}, std_{Angular}, mean_{CosSim}, std_{CosSim}\)). Calculating the euclidean distance is straightforward. Cosine similarity is commonly used in text analysis when comparing high dimensional bag-of-word vectors. It is calculated by:

\[
similarity = \frac{A \cdot B}{\|A\|\|B\|}
\]  

(7.9)

where \(A\) and \(B\) are vectors. Angular difference is calculated using similarity as follows:

\[
difference = \frac{\cos^{-1}\left(\frac{A \cdot B}{\|A\|\|B\|}\right)}{\pi}
\]  

(7.10)

**Bag-of-Visual-Words Data**

We have created a visual word dictionary of 1000 visual words from a database of over 3000 images ranging from photographs, to paintings, to digital images. With this dictionary, we build a 1000 element normalized bag-of-visual-words for the image being analyzed. We then summarize this vector with three statistics: the fraction of visual words from the dictionary contained in the image, the fraction of the vector taken up by the most common visual word in the image, and the variance of values in the vector (\(frac_{VW}, max_{Frac}_{VW}, var_{VW}\)).

We have also created a visual word dictionary of only 10 visual words from the same database of images. We build a normalized bag-of-visual-words for the image of interest with this dictionary. Each element in this vector is included as a unique feature (\(VW1, VW2, VW3, VW4, VW5, VW6, VW7, VW8, VW9, VW10\)).
7.4.4 Feature Selection

In total, we have described 198 potential global image features describing color (56), texture (50), shape (40), and local features (52). Since most machine learning algorithms learn most effectively with a few powerful features, we perform feature selection to identify the most useful features out of the 198.

For efficiency and consistency when extracting features, all images are scaled to a specific number of pixels along the largest dimension (thus maintaining aspect ratio). In this paper, all features contained in Table 7.2, (those obtained from DISCOVIR and the colorfulness features) were extracted from images scaled to 400 pixels along the largest dimension. All other features were extracted from images scaled to 800 pixels. The different scales are used because many of the DISCOVIR features are too costly to calculate at higher resolutions, yet the SLIP features are more effective with the higher resolution.

As our evaluation metric for feature selection, we use the average output separation of binary neural networks trained with backpropagation (or just separation for simplicity). Separation for a single binary neural network is calculated as the difference between the average output of positive and negative classes for that neural network. Separation is chosen because it correlates strongly with a variety of useful metrics—including Matthews Correlation Coefficient, inverse rank loss, and informedness—while being more consistent. In order to discover a set of features that can be used universally for affective image annotation, only the DARCI dataset—the most comprehensive of the datasets—is used for feature selection. Since the other three datasets are not used in feature selection, we can test the universal nature of the selected features with these datasets. We calculate separation as the average separation across all of the DARCI dataset categories.

We explore two feature selection algorithms, forward and backward selection. With forward selection, we begin with zero features and add one feature at a time—the feature that if added results in the highest separation for the classifier—until all features have been added. With backward selection, we begin with all features and remove one feature at a
time—the feature that if removed results in the highest separation for the classifier—until all features have been removed. Once the features have been ordered by each algorithm, we select only the top performing features for future use. Unfortunately, these two feature selection algorithms are computationally expensive as the number of features increases; so, we perform these algorithms on subsets of the features, then combine the improved subsets and repeat the procedure as indicated in Figures 7.5 and 7.6. The subsets were chosen to group similar features together that would most likely contain redundancies. For clarity, the fact that both selection algorithms yielded the same number of features is purely coincidental.

To determine the classifier’s separation at each step of the feature selection algorithms, for each category, we create 5-folds for cross validation. We use these same folds, and the same random seed for training (neural net initial weights still use random seeds) at each step of selection. These procedures are followed in order to properly determine the effect of adding or removing a feature by ensuring similar conditions at each step of the feature selection algorithms.

From each subset of features, we select the top performing features by ordering them with the two selection algorithms, and then picking the best features until there is essentially no significant gain in separation. To do this, we first approximate the amount of variance to

Figure 7.5: The flow of forward feature selection. Each box represents a subset of features, how many features the subset began with, and how many are selected from that subset.
Figure 7.6: The flow of backward feature selection. Each box represents a subset of features, how many features the subset began with, and how many are selected from that subset.

be expected between runs of the 5-folds for the same set of features. Since the 5-folds are consistent, and we use the same random seed for all datapoint permutations, this effectively tells us how much variance is caused by the initial random weighting of our artificial neural nets. We approximated this variance by running the 5-folds on the classifier 100 times using all 56 color features and taking the standard deviation of the resulting separation scores. This value, which we call $Std_{ANN}$, is calculated to be $5.412 \times 10^{-4}$. Next, we calculate the change in separation between the ordered features, $\Delta S$. Finally, starting with the best feature for each algorithm, we continue picking features until $\Delta S < Std_{ANN}$ for three successive features.

As an example of this process, see Figures 7.7 and 7.8. Figure 7.7 shows the separation scores for the SLIP subset of features using forward selection. Figure 7.8 shows the ordering of these features and indicates the point at which there is no significant gain in separation. The resulting features are then combined with the texture and shape features identified through forward selection for another round of forward selection.

In the interest of space, we don’t show the results of the selection algorithms on every subset of features. However, we do show the results of the final selection process for both forward (Figures 7.9 and 7.10) and backward (Figures 7.11 and 7.12) selection. We will refer
to these two sets of features as *forward* and *backward* features. Of note, 16 of our showcased SLIP features appear in each of the *forward* and *backward* sets.

In addition to these two feature sets, we will examine two other sets in this paper. The first, called *effective*, is the set of 39 features that is the intersection of the forward and backward features (Figure 7.13). 13 of these features are SLIP features. The second set, called *original*, is the 102 image features found in Table 7.2 that we have used in previous research.

### 7.5 Evaluation

For evaluation of our model and features, we use a metric that we call *true rate*. *True rate* (TR) is the average of the *true positive rate* (a.k.a. sensitivity or recall) and *true negative rate* (a.k.a. specificity). This metric is proportional to *informedness* which is commonly used in psychology and, along with DeltaP and Matthews correlation coefficient, has been shown to be less biased than other common measures including recall, precision, F-measure, and accuracy [76]. Formally, informedness is defined as: \( \text{true positive rate} + \text{true negative rate} - 1 \).
Figure 7.8: The ranked list of candidate SLIP features for forward selection. Those highlighted in blue are actually selected.

It follows then that TR (or *true rate*) can be converted into informedness by the following equation:

\[
\text{informedness} = 2 \cdot TR - 1
\]  

We use TR instead of informedness in order to allow direct comparison to Machajdik’s results [58]. While not completely clear, it is our understanding that when Machajdik indicates *average true positive rate* in their results, they are essentially computing TR by averaging the *true positive rate* of the positive and negative classes in one-vs-all classification for each emotion. By itself, *true positive rate* is a poor measure of success since it can be easily exploited by changing the threshold for classification without considering the number of negative instances that are correctly classified.

In all cases, except where noted otherwise, we perform 5-fold cross validation averaged over 50 iterations to obtain the displayed results. Since we are using binary classifiers, we need positive and negative datapoints for each emotion category. These are automatically provided with the *DARCI* dataset but must be inferred with the other three. To conform with Machajdik, we use the following one-vs-all approaches to create negative data for these
datasets. For the Art Photo dataset this is straightforward since each image is labeled with one emotion. For the IAPS dataset, we use the discrete labels established by Mikels [62]. Only the images not labeled with a particular emotion are negative datapoints for that emotion. For the Abstract dataset, for each image, the emotion with the most votes determines the label for the image. For all other emotions, the image is a negative datapoint. Any case where there is more than one emotion with the most votes is thrown out. As all three of these datasets use the same emotion categories, we create an additional dataset that simply combines all of the datapoints just described. This dataset is called Combined.

Since the DARCI dataset is so sparse, in order to have ample training and testing data for each adjective in the following experiments, we use only those adjectives with at least 30 distinct images that are positive examples and 30 that are negative examples. Thus, we will be looking at 110 adjective categories (see Table 7.1). For evaluating using the DARCI dataset, we compress all duplicate adjective labels for a given image into a single positive or negative datapoint. In the occasional event of conflicting labels (positive and negative labels for the same adjective-image pair), we determine the fraction of positive votes to negative votes and use that fraction as the sentiment of the datapoint rather than a binary 1 or 0. Depending on how the appropriate neural network classifies the datapoint, we will either
add this fraction to the true positive count or false negative count. The inverse fraction will either be added to the true negative count or false positive count. Thus these counts, used in calculating true rate, can contain non-integer values when using the DARCI dataset.

7.6 Results

First, we determine which set of features is best overall. While Machajdik [58] and Yanulevskaya [94] both select sets of features specialized for each emotion and dataset in their experiments, we are interested in a universal set of features that can be used across any adjective, including adjectives that our system has yet to encounter. For the remainder of our experiments, we will use only the best overall set of features that has been discovered using strictly the DARCI dataset.

Figure 7.14 shows the average true rate across all categories for each dataset for each set of features. Clearly, the new feature sets we have discovered in this paper out perform the
Figure 7.11: A graph indicating the separation score of neural nets trained using the indicated number of features. Removing features starts being detrimental once there are 51 features. The corresponding features are indicated in Figure 7.12.

original 102 features we have used in the past. This can be attributed to a thorough selection process and the addition of new features (more color features and the new SLIP features) since these new features comprise the majority of the selections. The biggest improvement can be seen with the DARCI dataset, likely because the features were selected using exclusively this dataset. Of the new feature sets, forward selection works best overall so we will use this in all further experiments. Effective selection is close, and contains fewer features, so it may have some performance benefits in future research.

7.6.1 IAPS Dataset Comparisons

Figure 7.15 shows the results of our neural network model trained with our set of forward selected features compared to the results obtained by Machajdik using their own features, Yanulevskaya’s features, and Wang’s features on the IAPS dataset as reported in [58]. Machajdik’s results using all three sets of features were obtained using a naïve Bayes classifier. They explored other classifiers including Support Vector Machines and the C4.5 algorithm, but got the best performance with naïve Bayes.

Our approach outperforms the other three in all categories except for Machajdik’s features in the case of amusement and Wang’s features in the case of sadness. For amusement,
Figure 7.12: The ranked list of candidate features for backward selection. Those highlighted in blue are actually selected.

Machajdik’s features outperform all others due to the use of three features specifically tailored to detect faces and skin in images. As Machajdik notes, the IAPS *amusement* images contain significantly more human faces than the other categories. This hints at the value of content-based image retrieval (CBIR) features in emotion detection; though, outside of face detection CBIR is a challenging prospect itself.

In Figure 7.16 we compare our approach directly to Yanulevskaya’s as reported in [94]. Here Yanulevskaya et al. use a Support Vector Machine trained with their own features and measure classification accuracy on the IAPS dataset. To conform to their methods, instead of performing 5-fold validation for these results, we train on 70% of the images in each category and test on the other 30% averaged over 50 iterations. Clearly, our model and features improve on theirs in every category, although it should be mentioned that accuracy is not the best metric of evaluation.
Figure 7.13: The unranked list of features that are selected by both forward and backward selection. We call these effective features.

Figure 7.14: The average true rate for all datasets for each set of features described in Feature Selection.

We do not compare our results directly to Wang et al.'s because they do not provide their datasets for direct comparison [92].

7.6.2 Art Photo Dataset Comparisons

Figure 7.17 shows our results compared to the results of Machajdik and Borth on the Art Photo dataset as reported in [58] and [9] respectively. Borth et al. use SentiBank, a library they have developed to detect adjective-noun pairs (ANP) in images [8]. SentiBank has been finely tuned to detect 1,200 ANPs using SVMs trained on an assortment of state-of-the-art visual features including (but not limited to) color histograms and a 1,000 element
Figure 7.15: The average performance of our approach (Norton) on the IAPS dataset compared to approaches using a naïve Bayes classifier with Machajdik’s, Yanulevskaya’s and Wang’s features as reported in [58].

Borth’s SentiBank approach performs the best in most categories, but overall, we have a comparable true rate score. Their success with SentiBank demonstrates the value of feature histograms which we summarize in our feature sets. Borth et al. also demonstrate the power of what is essentially an ensemble by learning semantic descriptors that in turn become features to another model. In future research, we may be able to use a similar technique by turning adjective descriptors into a feature vector.

### 7.6.3 Abstract Dataset Comparisons

In Figure 7.18 we see our results compared with those of Machajdik using their own features and Wang’s for the Abstract dataset as reported in [58]. As with the IAPS dataset, we do not compare our results directly with Wang et al.’s since they do not provide their datasets. In 6 of the categories, our model and features surpass the other two. The biggest exception is with anger. However, after preparing the Abstract dataset for binary classification, there are only
3 positive datapoints for anger. Naturally, these are not enough datapoints to appropriately train a neural network. The statistical nature of naïve Bayes clearly has an advantage in this case; although, if we were to negate the neural network output, it would outperform the naïve Bayes classifier.

### 7.6.4 Combined Dataset Comparisons

For the Combined dataset, we only compare our results to Machajdik’s as reported in [58]. As we see in Figure 7.19, our system outperforms theirs in 5 of the 8 categories. Again, we emphasize the fact that we use one universal set of features for all categories and datasets, while Machajdik et al. use the best features for each category and dataset. Furthermore, our set of features was selected without using any of the combined datasets (and different categories), demonstrating its generalizability. For the Combined dataset at the very least, it is likely that optimizing our features for each category and dataset would result in superior scores across the board.
7.6.5 DARCI Dataset

The 110 adjectives that we evaluate from the DARCI dataset can be seen in Figure 7.23(a). Recall that the adjectives listed actually refer to unique adjective synsets from WordNet. Since we are most interested in emotions, we have highlighted adjectives that arguably could be considered affective. We have also indicated the overall average true rate for all adjectives. There are 45 adjectives with a score greater than the average, 13 of which are affective. Referring back to Figure 7.14, we achieve better results on average with these affective words than any of the 8 emotions from the other datasets. Importantly, as the results of Borth et al. suggest, the detection of non-affective adjectives could be used in the future as features to detect emotions [9].

7.6.6 Revised Datasets

While the observed state-of-the art approaches to the affective annotation of images usually perform significantly better than random, they are far from satisfactory. Across all four datasets, plus the Combined dataset, our approach scores an average true rate of only 0.63—there is much room for improvement. That being said, we argue that the way many of these
datasets are handled is actually unfair to the evaluation of the approaches. The one-vs-all approach is common in emotion classification, but consider what this evaluation paradigm implies. In the case of the Art Photo dataset, an image that has been labeled with fear by its author is automatically considered not angry, not sad, and not disgusting. Can’t an image elicit any combination of those emotions—even all four? Of course it can; and yet, using the one-vs-all approach to generating negative classes penalizes a system for perhaps accurately detecting anger in a photograph that has only been labeled with fear. We contend that more consideration needs to be made when making assumptions about negative datapoints with these datasets.

Here we present the results of augmenting three of the datasets with simple, but more reasonable assumptions than those originally made.

First, we revise the Art Photo dataset by only creating negative datapoints for a specific emotion when images are labeled with an emotion of the opposite valence. For example, if an image was labeled amusement, it would only generate negative datapoints for anger, disgust, fear, and sadness. This is not a perfect assumption, but is more conservative than also generating negative datapoints for awe, contentment, and excitement for the same...
image. The results of using this revised dataset compared to the original one can be seen in Figure 7.20.

Next, we revise the *Abstract* dataset by using more information from the votes Marchajdik collected for each painting. Instead of only counting the emotion with the most votes for each image as a positive datapoint and all others as negative datapoints, we normalize the votes for each image and count every emotion with a normalized vote over 0.25 as a positive datapoint, and every emotion with a normalized vote less than 0.15 as a negative datapoint. The results can be seen in Figure 7.21.

Finally, we revise the *DARCI* dataset. This dataset benefits by not making any assumptions about negative datapoints. However, as we indicated earlier, it suffers from a scarcity of data for the many adjectives it contains. In particular, there are generally fewer negative datapoints than positive ones. Realistically, this fact should be reversed since there should be fewer adjectives that describe a particular image than adjectives that don’t describe the same image. Here we augment the *DARCI* dataset by adding negative datapoints for each positive datapoint by using the *related concepts* and *antonym* relationships in WordNet. Specifically, we add all WordNet antonyms for all WordNet related concepts for each positive
Figure 7.20: The average performance of our approach (Norton) on the original and revised Art Photo datasets.

datapoint. These added adjectives become negative labels for the image in question. For example, if an image were labeled positively with *happy* then we would add negative labels for *sad* and *gloomy*. Once we have augmented the dataset, we use the same 110 adjectives used previously for our experiments. The results are shown in Figure 7.23(b).

Figure 7.22 shows the overall impact of revising each of these datasets. Clearly, every one of these revisions improved the reported results. Furthermore, we would argue that these results are a more accurate assessment of our classifiers and features. Arguably, more sophisticated revisions could be made to provide even more accurate (and probably improved) results. We encourage future researchers to make such considerations when evaluating ESIR systems.

### 7.7 Conclusions

We have presented several new image features and shown their value in the annotation of images with emotions across four datasets. Three of these datasets are well established, and we compare our results with four state-of-the-art approaches across these datasets. In particular, we have demonstrated the strength of SLIP (Summary of Local Interest Point)
features in combination with more traditional color features. The features we have introduced in this paper will contribute to the growing wealth of image features used in CBIR research.

We have also introduced a new dataset for the ESIR community—the DARCI dataset available online. This dataset contains numerous unambiguous adjective labels compatible with WordNet, as well as explicit negative labels for these adjectives. The dataset also includes over 2000 images from varied mediums (paintings, photographs, digital art, etc.). We have shown that our neural network model with our set of features can adequately label images with many of these adjectives. In the future, we hope that researchers in the ESIR community will continue to make use of the datasets provided by Machajdik et al. and Mikels et al. to allow for direct comparison between methods. Likewise, we hope that our DARCI dataset will be utilized for its benefits.

Finally, we have presented arguments against strict one-vs-all approaches to evaluation in ESIR research. We have provided a few alternative methods of obtaining implicit negative data that make fewer extreme assumptions about the datasets.

In future research, we will explore the augmentation of the DARCI dataset through natural language processing techniques that take advantage of the semantic relationships of adjective labels. Furthermore, we will explore the possibility of turning adjective labels into
new feature vectors for more accurate emotion detection. Finally, we will incorporate the new feature sets we have discovered here directly into our creative system, DARCI, to further our work in the field of computational creativity.

Figure 7.22: The average performance of our approach (Norton) on original and revised datasets.
Figure 7.23: The true rate of 110 adjectives from the (a) DARCI dataset and (b) revised DARCI dataset. All results are obtained using binary neural networks and our forward selected features. Affective adjectives are highlighted in green; the average true rate of all 110 adjectives is in red.
Chapter 8

An Artistic Dialogue with the Artificial

Abstract

In conjunction with Brigham Young University’s Visual Arts program, we conducted a study centered around a system designed to be an artificial artist, in order to synthesize the ideas of visual artists and computer scientists. Participants from both disciplines designed activities that imposed the limitations of the artificial system on their fellow participants. These activities sparked discussion and insight into the nature of the creative process and how it can be better emulated in artificial systems. We present our system and several of the activities designed around it and discuss the synergistic results.

\footnote{David Norton, Derrall Heath, and Dan Ventura. An artistic dialogue with the artificial. In *Proceedings of the 8th Association for Computing Machinery Conference on Creativity and Cognition*, pages 31-40, 2011.}
8.1 Introduction

In order to instigate a dialogue, a group of college art students was presented with a series of anonymous images and asked to comment on the creativity and value of the images. The images ranged from Van Gogh’s *Starry Night* to Malevich’s *Red Square* to a fourteen year old student’s meticulous reproduction of *Starry Night*. They included a painting titled *Purple Haze* by Rumba the Wonder Horse, a digital image from the Mandelbrot set, and a phase contrast micrograph of *Trichodina pediculus*. They also included three images each produced by one of the digital artists: AARON [61], The Painting Fool [13]2, and DARCI [68]. The students knew the work of Van Gogh and Malevich. They could immediately spot the fraud of the *Starry Night* reproduction. They marveled at the mathematical beauty of the Mandelbrot set and phase contrast micrograph. None of them knew the works of Rumba the Wonder Horse, or the three digital artists—while intrigued, they did not know how to appreciate these works. From these students’ perspective, they could not properly attribute creativity without understanding the process behind the creation.

This exercise inaugurated a four month study designed to create synergism between the (apparently disparate) disciplines. With the collaboration of students and faculty from Brigham Young University’s Visual Arts department, we designed the study to explore the process of creation in the visual arts. Our interest in the collaboration was to learn how to better emulate the human process of creation in our digital artist, DARCI. For the participating art students, it was an opportunity to explore their own personal approaches to creation and why they do what they do when they create.

The centerpiece of the study was DARCI (Digital ARtist Communicating Intention) a computer system designed to create original images through processes perceived as creative. We facilitate the perception of creativity by giving the system a name and referring to DARCI as a she. *She* is central to our research in the budding field of computational creativity. Computational creativity is a branch of artificial intelligence focusing on emulating human

2http://thepaintingfool.com/
creativity in computer systems. One of the goals of this field is to automate tasks that one would say require creativity to perform such as mathematical deduction, the creation of art and music, and storytelling. Another goal of the field is to better understand creativity so that we may identify, foster, and augment the creative process in society.

DARCI presently does not produce wholly original images; rather, using a variety of image processing filters, she renders provided images so that they will reflect an accompanying description, currently limited to adjectives (see Figure 8.1 for example). In accordance with Colton’s arguments concerning appreciation [12], this process requires that DARCI be able to evaluate the degree to which an image matches a list of adjectives. She is able to make this assessment by learning to associate image features with individual adjectives via supervised machine learning. In other words, the data used to train DARCI is created by teachers—humans. The opportunity to teach DARCI is open to the public with no restrictions through a persistent public website. What this amounts to, is that DARCI’s interpretation of images is an unpredictable amalgamation of human sentiment and the machine learning algorithms at play behind the curtain. Effectively, DARCI presents a unique interpretation of the social culture surrounding her development—arguably the role of an artist.

Clearly, DARCI is limited by constraints far more restrictive than her human counterparts; at the same time, these constraints provide a unique viewpoint. During the study, we engaged in activities designed to explore the constraints of DARCI from an artistic perspective.
in an attempt to stimulate introspection about the process of creating art in the participants while helping us improve DARCI. The study culminated in *Fitness Function*, an interactive art show, hosted by Brigham Young University’s Visual Arts Department, in which DARCI acted as the sole juror for any-and-all submissions made on site.

This paper will explore the insights gained from simulating current computationally creative systems with social activities. We will begin by detailing the processes by which DARCI currently learns and creates, and observe the constraints imposed by these processes. We will then describe the activities, designed by both artists and computer scientists, that were employed during the study to foster an understanding of DARCI’s limitations and provoke discussions about creativity. We will illustrate how these activities sparked insights into the creative process by those participating in the study. We will also outline how these insights will be used to redesign DARCI and improve her capacity to be perceived as a creative agent. Finally, we will describe the art show, *Fitness Function*, and the response it received from those who participated in it.

### 8.2 DARCI

An important component of creativity is appreciation, or the ability to evaluate one’s own created artifacts [12]. In order for DARCI to appreciate art, she must first acquire some basic understanding of art. For example, in order for DARCI to appreciate an image that is bright and cold, DARCI must first understand the concepts ‘bright’ and ‘cold’. To do this, DARCI must learn to associate images with descriptions. We teach DARCI to associate images with descriptions by providing her with hundreds of images, each labeled with adjectives that describe it. We also provide negative labels, or adjectives that the images are not. DARCI considers low level features of each image that deal with light, color, texture and shape. These image features are then associated with the provided adjectives through a collection of artificial neural networks [66]. The more examples of happy and non-happy images that are provided, the better DARCI is able to understand what it means for an image to be
happy and be able to evaluate how happy an image is. We collect labeled images for DARCI through a website that allows people to label images with adjectives (http://darci.cs.byu.edu). The user is presented with a random image from DARCI’s database and the user is asked to provide adjectives that describe that image. Additionally, there is a section of the website in which DARCI attempts to label an image with adjectives herself and the user can correct DARCI by tagging the adjectives that are incorrect, thus providing negative labels.

Once she has sufficiently learned enough adjectives, the next step for DARCI is to modify an image to match a list of adjectives. DARCI has a large set of image filters, similar to what would be found in software like Photoshop, which are capable of modifying images. DARCI uses genetic algorithms to explore different image filter combinations and to learn how to use these filters to modify images to convey the meaning of each adjective [68]. Genetic algorithms have often been described as a creative process due to their purposeful, yet unpredictable nature [33].

Genetic algorithms are governed by a subroutine called a *fitness function* that evaluates each artifact. The fitness function for DARCI generates a single score that measures how well a set of filters modifies an image to match the provided adjectives. This fitness function is the output of the artificial neural networks that have been learning image-to-adjective associations. For example, if DARCI were given a photograph and the adjectives ‘bright’ and ‘happy’, then DARCI would begin the creation process by trying out random sets of image filters on the photograph. For each set of image filters, DARCI would score how ‘happy’ and ‘bright’ the resulting image is using the artificial neural networks. The sets of filters with the lowest scores would be discarded, while DARCI would keep the sets of filters with the highest scores. DARCI would then alter some of these filters (*mutation*) and even try recombining the best sets of filters (*crossover*). In this way DARCI is continually searching for the set of filters that would modify the photograph to be the most ‘happy’ and ‘bright’. This process continues for many iterations, after which DARCI applies the best set of image filters found and outputs the modified image. A high level overview of DARCI can be seen in Figure 8.2.
DARCI is a simplified model of a complex process and hence has several important limitations. For example, DARCI's language is limited to lists of adjectives. Her percepts are limited to low-level image features and her ability to modify images is limited to a finite set of image filters. DARCI has no social context or past experience from which to draw analogies and insights. Finally, DARCI's feedback consists of a single metric with no context. These limitations will be discussed in greater detail throughout the paper.

8.3 Artist Interpretations

We began the study by introducing the participants to DARCI. Soon after that, still early in the study, we asked several of the artists to come up with activities for the rest of the participants that they felt represented the processes underlying DARCI as they understood her. This was not only to solidify their understanding of DARCI, but also to help the artists begin to think about their own processes of creating art. The artists came up with activities that highlighted many of DARCI’s aforementioned limitations. This section contains a description and analysis of three of the activities that the artists devised.
8.3.1 Copy of a Copy

Copy of a copy was based on the simple premise that as information is transferred, it degrades. The participants were divided into four groups. Each group created a black-and-white picture as a team using only pencil and ink. There were no other rules. Once each team had finished their pictures, they found a copy machine on campus and made a copy of the picture. They then made a copy of the copy. They proceeded to copy each copy for fifty iterations. The groups then convened to show their results and discuss what happened.

As expected, the quality of each copy deteriorated to the point that the original image could no longer be recognized. Some groups experimented with different patterns or shade strengths. While the fidelity of each copy decreased, interesting artifacts would sometimes occur due to the way the original images were drawn. The participants decided that from an artistic perspective the degradation, to a point, was not always a negative thing.

This activity was designed by an individual who was concerned that as DARCI was taught adjectives, the quality of her learning would be suspect. Said another way, since both DARCI’s percepts and the communication between teacher and DARCI are not perfect, there will always be a degradation of information. This individual assumed that DARCI was meant to reflect the opinion of her teachers. After this activity, the general consensus was that this imperfection in learning was, in fact, a good thing and not unlike a human teacher-student interaction. The flaws in communication are part of what allow the novel to emerge. This idea was explored throughout the study and the many activities we participated in; but this activity set the tone.

8.3.2 Rotating Artists

This activity had participants creating collaborative drawings all using the same medium, pencil. An 18 x 24 inch sheet of paper was set up on a drawing board for each individual. Each sheet was associated with a random adjective. Then each person was assigned a random rule restricting the mechanics of their drawing. These rules included requirements such as
“draw with your off hand”, “draw from over the top of the paper”, “draw with your teeth”, etc. The participants were randomly assigned a sheet of paper to start on. They were instructed to begin a drawing that fit the adjective assigned to the sheet of paper. After five minutes, the participants rotated to another sheet of paper. They were instructed to continue the previous individual’s drawing using the adjective connected to the sheet. Participants continued to rotate until everyone had contributed to every drawing. Individuals were required to use their restriction on every other rotation. Finally, when all of the drawings were completed, the participants were split into three groups and each group created a collage out of one third of the drawings.

This activity simulated DARCI on several levels. First, participants were given a task similar to DARCI’s primary function: they were required to create an image to fit an adjective. Second, participants were given hardware restrictions, their drawing rule and use of only pencils. This was an attempt to constrain them in the way that DARCI is constrained to digital media using only filters. The restriction is not identical but it did force individuals to create without their complete faculties, giving them something to think about. Finally, there were two attempts to specifically simulate the way DARCI generates her artifacts. The first was in rotating people from drawing to drawing. The second was in cutting up the drawings and combining them in some form of collage. These activities were attempts to copy the crossover and mutation mechanisms used in the genetic algorithm driving DARCI’s image creation. Mutation is an iterative mechanism not unlike the incremental alterations contributed to each drawing by new artists. Crossover does combine features from multiple solutions as the collages did. However, unlike actual genetic algorithms, there was no evaluation of the changes made at each step and no decision based on these evaluations as to how to proceed. Also, the goal of the activity differed from DARCI’s in that the participants were creating a wholly original image rather than modifying a source.
This was the first activity of the study to directly simulate DARCI’s processes. As such, it began to bridge the gap between the understanding of the artists and that of the computer scientists.

8.3.3 Systems

In this activity the designer attempted to more accurately simulate DARCI’s genetic algorithm based on the discussion deconstructing Rotating Artists. In addition, the designer added some style by presenting the activity as if those participating were themselves machines. Systems was divided into two sessions called System 1 and System 2 respectively. Each participant was given a unique Ink Hardware, a tool for applying ink on paper (i.e. various pens, ink blots, an eye dropper, etc.). The participants were then presented with the instructions shown in Figure 8.3 and each given a card similar to the one in the instructions. Basically, within sixteen minutes, each participant created an image of the noun listed under Program using their “hardware”. They then wrote a word that described their experience under Source Code. Finally, they paired up with another random participant and performed the System Purge to determine which individual would progress in System 1 and which would be eliminated and move on to System 2. Those that progressed within System 1 got a new card with a Program to modify the eliminated participant’s art in some way. Individuals continued to pair up and System Purge until there was only one person left who then created the final image.

The results of the four rounds of elimination that occurred in System 1 are shown in Figures 8.4-8.7(a) respectively. The final image is shown in Figure 8.7(b). This activity simulated the generational progression of DARCI’s genetic algorithm. Each generation, the “optimal” images would move on to compete in the next generation. Furthermore, each generation was built off of the previous one. While an analogue for crossover was not present, one for mutation was. The significant difference between System 1 and DARCI’s genetic algorithm was that the criteria for success in an image was not directly correlated with the quality of the image. The winner of each System Purge was determined by either winning
Figure 8.3: The instructions for System 1 as presented to participants.

a thumb war, having the fewest siblings, being born furthest from the activity location, or wearing the most articles of clothing, depending on the round. Typically, such an indirect evaluation criteria is not ideal in genetic algorithms. However, it can have unforeseen and fascinating implications (maybe people with strong thumbs make better artists).

As participants were eliminated, they moved on to System 2 (see Figure 8.8). System 2 was essentially a group project where each panel iterated on the previous one. People were assigned a panel to complete in the order they were eliminated from System 1 and then by birth date for individuals eliminated simultaneously. Each panel was supposed to reflect the predetermined Program and the previous panel. Participants could use whatever Ink
Hardware was available at the time they started their panel. This system explored the concept of evolution at a high level. Mutation occurred in the sense that individuals iterated upon one another’s work. The progress of the panels was aided by an evolutionary theme in the Program titles. The Program themes progressed in the following order: birth, adaptation, bewilderment, fear of other, conflict, power, unification, stress, pain, relief, disbelief. The result of System 2 can be seen in Figure 8.9. While open to interpretation, the Program themes hint at a message about the process of merging ostensibly conflicting viewpoints: say that of visual arts and that of computer science. With this in mind, Systems not only demonstrated an artistic interpretation of DARCI’s creative process, but also acted as a metaphor for the study as a whole.
8.4 Computer Scientist Interpretations

When an artist creates art, what is going on in their head? Said another way, given the artist’s inputs, what determines the output? In this context, the input includes feelings, education, training, genetics, the weather, day-to-day senses—everything that one experiences. The output is the art. Some process occurs that transforms these inputs into artifacts. The activities that we designed were aimed at trying to understand this process. The first activity, Bromerly, was aimed at exploring the process of learning associations and expressing them in
art. The second, *Celebrities Crossover*, was aimed at exploring the idea that meaning is not universal. The third activity, *External Evaluation*, was aimed at exploring the evaluation of art from both the creator’s and the critic’s perspectives.

### 8.4.1 Bromerly

When we, as humans, hear adjectives like ‘happy’, ‘scary’, or ‘red’, we have a lifetime of experiences and examples to help us understand what these words mean. These adjectives fit into a cultural and environmental context that influences how we interpret them. DARCI has none of these advantages. When DARCI learns a new adjective like ‘gloomy’, she has no past experience to draw upon, and she has no cultural influences to help define it. All DARCI has are the example images of ‘gloomy’ and ‘not gloomy’ that are given to her.

An activity was designed for the human artists to simulate DARCI’s processes and limitations as much as possible. The goal was to see how they would compare to DARCI, to determine what their creative processes would be like in a restricted setting, and to discover improvements for DARCI. Three fake adjectives were invented that have no reference to

![Fourth round results](image1)

![Final image](image2)

Figure 8.7: Fourth round and final image results for *System 1*.
any other word, and have no background or cultural meaning. These adjectives are ‘orfly’, ‘bromerly’, and ‘flamping’. Unbeknownst to the human artists in the study, these adjectives actually represented a real adjective that DARCI knew fairly well. Orfly meant weird, bromerly meant blocky, and flamping meant lonely. The artists were then provided with the same example images for each adjective that DARCI was trained on (both positive and negative examples).

The participating artists were split into three groups. Each group was given one of the fake adjectives and the example images for that adjective. The artists spent several minutes looking over the provided images to learn as best they could what the fake adjectives meant. Each artist was then given their own source image and instructed to create something based off the source image that reflected the adjective that they just learned. Essentially they had to communicate the meaning of their adjective through their creation. Some of the resulting images along with their sources can be seen in Figure 8.10. After everyone was done, all the paintings were put together. A random group of college students, not involved with the study, was then brought in to divide the paintings into three groups. These people had no
knowledge of the activity or how they were supposed to group the paintings. The volunteers were able to group the paintings of the same adjectives together with the exception of one painting. The grouping results can be seen in Figures 8.11-8.13.

Although unable to specifically define their adjective, the artists were able to identify certain image features or qualities that they associated with their adjective. For example, they were able to tell that bromerly (blocky) images tended to have hard geometric lines, tended to be very structural, and tended to have a lot of divided space. They could tell that orfly (weird) images often had exaggerated curvy shapes and had some distinct peculiarity.
Figure 8.10: Three example images (with their sources) created during the Bromerly activity. One example of each fake adjective is presented. Flamping means lonely, orfly means weird, and bromerly means blocky.

They could also tell that flamping (lonely) images were usually dark and had a singular focus. The artists were then able to incorporate these characteristics into their own paintings. During the grouping phase, the people doing the grouping were able to pick up on these features and successfully group similar paintings together, while only miss-classifying one painting. This notion of evaluating the paintings by grouping led to a paper currently in submission that explores the use of clustering algorithms to evaluate DARCI’s artifacts.

Figure 8.11: The cluster of flamping (lonely) images.
A notable difference between DARCI and the human artists was how quickly humans were able to learn the adjective with so few example images. This is not surprising considering how sophisticated human vision is. DARCI needs hundreds of example images and can only see explicitly specified low level image features, while the human artists were able to quickly identify and extract the useful image features and then convey those features into their own paintings. Taking note of the way the human artists examined the example images and identified the important image features has provided us with useful insights into the type of image features DARCI needs to extract.

Through this activity, the artists gained a better understanding of how DARCI learns adjectives with the limitations that she has. The artists found it very interesting to learn a new word with only images rather than linguistically. They mentioned that when most words are heard, an associated image usually comes into people’s minds. For example, if someone hears the word ‘chair’, an image of a chair pops into their mind. It may be a different type of chair for each person, but some kind of image pops into their head.
“When I say chair, everybody thinks of a chair in their mind. They see that image, it might be a different chair. I just think that the images are the words.”

They also said that using visuals to convey the mean of a word is less ambiguous than using other words to explain the meaning because those other words have different meaning and connotations for different people, while images are more concrete.

“It is easier for us to get it visually than through words, words are very ambiguous. ... I usually have to see something visually to remember it.”

In order to provide a more direct comparison with DARCI, several artists were asked to learn additional fake adjectives and then, using only image filters found in Photoshop, modify a source image to reflect multiple adjectives. Restricting the artists to Photoshop ensured that they were limited to the same tools that DARCI has to modify an image. We had DARCI do the same task with the same adjectives and same source images. The results can be seen in Figures 8.14-8.16. The artists were again able to quickly identify the distinctive image features that they associated with their adjectives. Some artists were unfamiliar with Photoshop and spent time trying different image filters to learn how they would affect the source image. They would then find the ones that modified the source image to best reflect their understanding of the adjectives. This process that they described is very similar to the genetic algorithm process DARCI uses to learn the filters to modify images. Additionally, these results are consistent with prior research that shows that DARCI is capable of producing images comparable to human artists under similar constraints [68].

8.4.2 Celebrities Crossover

This activity was structured around the party game Celebrities. In this game, players secretly write the names of several people or characters (fictional or not, famous or not, made up on the spot or not) on about 40 pieces of paper. The players are divided into two teams. Each team takes 30 second turns by having one of its members describe the characters on
Figure 8.14: Images produced by a human and DARCI from the same source image for the combination of the adjectives bromerly (blocky), blemistic (majestic), and diambic (avian).

Figure 8.15: Images produced by a human and DARCI from the same source image for the combination of the adjectives flamping (lonely) and orfly (weird).

the pieces of paper while the other members try to guess the character. Every correct guess scores the team a point. This continues until all of the characters have been guessed. This process is repeated for two more rounds using the same set of characters and same teams but with different rules. During the second round, players can only use a single word to describe the character. On the third and final round, players can only use gestures to describe the characters. Since the same characters are used every round, as the game progresses, players develop unique associations to aid in describing these characters. Often these associations have nothing to do with a character at all, but stem from players attempting to describe characters they don’t know as fast as possible. These frequently bizarre associations get
picked up by everyone in the game (both teams) and can even become memes within the group that played the game together.

The study participants were divided into two groups (with two teams in each group). Each group played *Celebrities* with the exact same character names, but physically separated from each other. Thus, two divergent sets of character associations were allowed to develop. Once the groups had completed their games, one team from each group was transplanted into the other group. Each of the resulting groups then played a revised version of *Celebrities*. The new version of the game was no longer a team game; instead, it was everyone for themselves. The participant with the most points from each group would be the winner. The rules for the game were the same as round two from *Celebrities* (meaning that participants could only say a single word or use gestures) except that each participant was performing for everyone else in the group instead of teammates. Each individual that correctly guessed a character was awarded a point; the person performing only received points if the player who correctly guessed the character came from a different group than their original one. If this happened, the performing player was awarded 3 points. This forced the participants to deliberately avoid the associations they had formed previously, while developing new ones that would give everyone playing the same chance of guessing correctly.
In semiotics, the association between a sign and that to which the sign refers denotes meaning. An important aspect of visual arts is the ability of the artist to convey meaning to an audience through their art. Thus, the image that an artist produces would be the sign that is conveying the meaning that the artist intends. We designed this activity to stress the idea that not every culture associates signs with the same meanings. In this context, the term culture can refer to social groups of any scope, including an individual. When an artist creates a piece of work, to what extent is the artist communicating meaning? Is the artist only communicating to those that share the same associations as intended in an artifact? Do artists make an effort to communicate outside their personal semiotic domains? We intended for this activity to force those participating to ask themselves these questions. In answering these questions, we hoped to learn the extent to which the communication of meaning would need to be emulated in DARCI for her to be accepted as an artist.

In *Celebrities*, the signs are the gestures or single words, and the meanings are the character names. With DARCI, the signs are the feature vectors that she extracts from an image, and the meanings are adjectives. Since the feature vectors represent DARCI’s percepts of an image, we suggest that DARCI’s signs are effectively the images themselves. In *Celebrities Crossover*, each group forms its own semiotic domain for the set of character names. When the groups cross-pollinate, they are required to communicate outside these newly formed semiotic domains. In our experience, this involved relying upon alternative associations that existed prior to the activity—in other words experience outside the activity’s restricted gesture/word based symbolism. One shortcoming of DARCI that was brought to our attention by these observations was that currently, DARCI has no associations besides the image-adjective relationship. This means that she is unable to use experiences external to these associations to communicate with an audience outside her personal semiotic domain. Another shortcoming we discovered was that, even within the image-adjective relationships, DARCI is only able to model a single semiotic domain. This domain represents a function of all of the opinions of people who have trained DARCI weighted by how much time they
have spent with her. While the associations this creates may be interesting, there is no attempt to cater to a specific audience. We are redesigning DARCI so that she will be able to model each individual that trains her in addition to a general model. This will allow her to cater to specific individuals when commissioned—essentially communicating within that person’s semiotic domain for image-adjective associations. Whether DARCI should cater to individuals is another question; but, at least having the capability to communicate more effectively is something we are interested in.

Referring to *Celebrities Crossover*, the following comment was made by a participating artist:

“I was so fascinated by the associations that we made. Dan said he didn’t think it was a big deal that we made “incorrect” (insane, irrational...) associations, but does DARCI make associations that are incorrect? Does anything she do come out of random associations or is all of her information supplied by us? What happens if the associations she makes are incorrect? Maybe nothing. Maybe it doesn’t matter.”

Another artist replied with:

“The associations being created are not always expected. They may not be the prophesied outcomes. But that does not make them incorrect. Each association has been built upon an individual’s experiences. So for that individual the association is alive and correct. Even if it is not the association that was predicted.”

This discussion illustrates some ideas that were brought to the participant’s attention as they pondered the activity. Whether or not these ideas are actually valid in the domain of visual arts is irrelevant. What’s important is that the participants are thinking about the process of creation—their process of creation.

### 8.4.3 Abstracted Feedback

As described previously, DARCI uses genetic algorithms to learn what set of image filters will modify an image to reflect a certain adjective. This genetic algorithm uses a fitness function
to guide the exploration. For example, if DARCI was trying to make an image fiery, then DARCI would try several sets of image filters. The modified images would then each get a numerical score from the fitness function. This score tells DARCI how fiery each image is. This is just a single number—it does not tell DARCI anything about how or why it is or isn't fiery. All DARCI knows is that the higher the number the better. DARCI can then take the highest scoring sets of image filters and perform crossover and mutation on them to search for even higher scoring sets of image filters.

An activity was designed to give human artists this same limitation in feedback in order to see how they would perform, and to see what their thought processes would be. Each person was given a quarter sheet of paper and had to draw a picture that matched a specific criterion. To start with, no one had any idea what that criterion was. The only direction they were given was that their drawing would be scored based on some criterion unknown to them and they had to try to maximize their score. After each person submitted a drawing, the drawings were each given a numerical score. The drawings were handed back with their score, and each person could view the returned drawings and try to figure out why one drawing scored higher than another. Each person then submitted another drawing in order to increase their score. This process was repeated for several iterations.

The participating artists were split into two groups each with different criteria. The first group had a purposefully subjective criterion for which they had to draw a scary image. The drawings were each given a score based subjectively on how scary the human evaluators thought the drawing was. The second group’s criterion was more objective. They had to draw a specific shape with the correct size and location on the paper. Human evaluators scored each drawing on how well it matched a specific template (seen in Figure 8.17) by holding the template up to the drawing and counting how many square regions on the template overlapped with the drawing. Example drawings from group 1 and group 2 can be seen in Figures 8.18 and 8.19 respectively.
The artists got frustrated quickly when they would put time and effort into a drawing and then receive a low score and not know why. The first group at first submitted a very wide range of drawings. Eventually, someone drew a pirate ship which bumped up their score considerably. The others in the group keyed in on this higher score and all started drawing pirate related things. Soon every artist in group 1 converged to drawings of skulls. The second group was never able to get the shape completely right, they converged to a drawing of a shaded square. They did, however, converge roughly to the correct size and location on the paper.

This activity proved to be difficult and frustrating for the artists. To successfully submit a high-scoring drawing required a lot of time, persistence, and patience. This activity helped the artists understand how DARCI has to deal with limited feedback. It quickly generated a discussion on how art is evaluated in the real world. Many of the artists who had submitted their artwork to various art exhibits admitted that they were usually oblivious to the requirements of a particular art show. They would just submit all of their best work and hope something got in. This discussion developed into the idea for the art exhibit *Fitness Function*. 

Figure 8.17: The template used to evaluate the objective group.
Figure 8.18: Examples of drawings created by the group evaluated subjectively based on how scary their images were. Drawings are chronologically ordered from top left to bottom right.

8.5 Fitness Function

Fitness Function is an interactive art exhibit where DARCI acts as the sole juror for pieces submitted throughout the show. The exhibit begins with an empty wall that gets filled with art that DARCI selects from submitted work. DARCI is set up in the exhibit with a simple interface for submitting images via USB sticks. Participating patrons provide their name, email, the title of the piece, and the image itself and then DARCI scores the image between 0 and 100 using the neural networks mentioned previously. Images that receive a score of 70 or higher are printed for the artist to hang on the wall, and images that receive a score of 90 or higher receive a jury award. Artists can submit as many images as they want and anyone is eligible.

With the collaboration of Brigham Young University’s Visual Arts department, we set up Fitness Function in the Harris Fine Arts Center to operate between March 19th and 30th, 2010, with a closing reception on the 31st. Advertisements for the exhibit were spread throughout campus via school website, posters, fliers, and word of mouth.
We used a model of DARCI trained exclusively by the participants in our study. Just as potential images created by DARCI’s own genetic algorithms are evaluated based on how closely they match particular adjectives (described earlier), the images submitted by participants of *Fitness Function* were evaluated based on how closely their images matched a set of secret adjectives that we chose for the exhibit. The adjectives we chose were ‘phantasmagoric’ and a special artificial adjective that we created for the purpose of *Fitness Function*. The special adjective was DARCI’s prediction of whether a piece would be accepted to Brigham Young University’s Museum of Art. In other words, the adjective was ‘museum-of-artsy’. So that the artists involved in our study could participate, the adjective ‘phantasmagoric’ was kept secret from them; they weren’t even told about the existence of ‘museum-of-artsy’.

By the end of the exhibit, 93 people had submitted 787 images. 145 of these images were accepted, and 10 of those received a jury award at the closing reception. The jury
award winners are shown in Figure 8.20. A website documenting the show can be found at http://darci.cs.byu.edu/fitnessfunction/.

As indicated by the comment book in the exhibit, word of mouth, personal observations, and the nature of submissions, people’s reactions to the exhibit varied significantly. Some felt pride and vindication due to the acceptance of their art: “I MADE IT IN! YEAH!”. Some felt frustration: “Fun and frustrating at the same time. More like a game to try and figure out than serious.” Others took it as a personal challenge. One individual submitted 106 images, clearly experimenting with lighting, image quality, and other less conventional alterations (Figure 8.21). Thirty-eight of his images were accepted. As the last entry in the comment book he wrote, “IT WAS FUN WHILE IT LASTED DARCI.” Whatever their opinion, people couldn’t help but personify DARCI: “An interesting project, however, it seems that DARCI is very opinionated in her ratings. From a look at the wall, she has some rather odd tastes.”

In this exhibit DARCI represents a metaphor for all kinds of external evaluation criteria to which an artist is subject, whether implicitly, explicitly, consciously, or unconsciously. We wanted to challenge people’s notions of evaluation. How does one decide that they have succeeded with a piece of art? Does one only esteem a work if it is valued by others; if it can
Figure 8.21: Two sample submissions from a witty artist.

get into an art show? What if one’s art gets a high score from a machine? That was quite significant for some people. The ability to evaluate artifacts on some level is critical in any act of creativity, and from this exhibit we learned that people at least perceive that DARCI has an opinion (i.e. that DARCI is capable of evaluation). The perception of creativity has been argued to be creativity [12]. If this is the case, then shows like Fitness Function are a step in the right direction for indentifying arguably creative systems.

8.6 Conclusions

During the study, the question was raised: “does DARCI think?” The computer scientists responded with “if we can describe it, we can program a computer to do it.” This quote encapsulates the computer science viewpoint and our motivation for the study. If we can formalize the process of thinking, then a computer can be programmed to think. If we can formalize what it means to be creative, then a computer can be programmed to be creative. The activities in the study were designed to elicit personal introspection on one’s creative process in the hopes of shining light on that elusive process. By focusing the activities on DARCI’s limitations, we were able to help those participating in the study isolate the thoughts and actions behind their creations.

As illustrated in the previous sections, we were able to gain some insight for our system, DARCI, through the activities we designed and the discussions that followed: the degradation of information associated with learning is acceptable, clustering algorithms can be used in
the evaluation of artifacts, more realistic image feature extraction is vital, the creation of perceptual associations is important (perhaps in all creative systems), genetic algorithms do appear to emulate part of the creative process, the creation of unique relationships is acceptable and useful, being able to create relationships across multiple senses is necessary to communicate across semiotic domains, and finally the external evaluation of one’s work is significant in the creation process.

Throughout the study, participants were encouraged to contribute to a blog titled Art[ificial] (http://computationalcreativity.blogspot.com/). This blog was designed for participants to leave their thoughts on the activities and discussions we engaged in during the study. Consider the following entry:

“Today I was thinking of interpretation. I interpret a lot of things. I think that’s a good word to put to my process of taking a lot of input and making an output that I feel will have significance in a realm beyond my personal self. I am interpreting emotions, thought processes/concepts, words, into a visual language. Other words I think might be synonymous are ‘translating,’ or ‘encoding.’ Often times I have heard the word ‘feel’ used in critiques, or have thought it to myself, where something simply feels consistent, or right, or something feels like it doesn’t belong. It is a way of interpreting things across different areas, from visual to emotional, or visual to psychological—a synesthesia of emotion, subconscious, and physical senses.

I read some interesting articles about synesthesia and creative processes, and I think that is something significant to art making—the ability to have ideas cross over from one realm to another, or one language to another, while still maintaining a certain degree of integrity or self sufficiency.”

This thought sums up the nature and purpose of the study—the synthesis of artistic and scientific interpretations for the goal of understanding the process of creation.
8.7 Acknowledgments

We thank Professor Joseph Ostraff and his VASTU 480/680 students for the fruitful collaboration.

This material is based upon work supported by the National Science Foundation under Grant No. IIS-0856089.
Chapter 9

Exhibitions

Throughout the development of DARCI, we participated in several art exhibits and galleries. DARCI’s involvement in these exhibitions ranged from curator, to artist, to collaborator. In this chapter, we summarize each of these events.

9.1 Fitness Function

The collaboration with Brigham Young University’s Visual Arts program, described in Chapter 8, culminated in an interactive art exhibit held in the Harris Fine Arts Center at BYU between March 19th and the 30th, 2010. The exhibit, called Fitness Function\(^1\), began as a room with only DARCI (installed in a computer), a printer, some instructions, and an empty wall. Visitors submitted their own digital art and DARCI evaluated it in the same way the system evaluates its own artifacts. Those submissions that achieved a sufficiently high score were printed and the visitor could put their piece on the wall. The exhibit gradually grew into a full gallery curated by DARCI (see Figure 9.1). Fitness Function elicited a variety of responses from visitors, challenging many of their perceptions about art curation. This collaboration illustrates how DARCI has influenced artists as a source of inspiration and introspection.

Fitness Function was so successful that we repeated the exhibit for one night during the Conference on Creativity and Cognition in Atlanta, GA. The second Fitness Function

\(^1\)A website documenting Fitness Function can be found at: http://darci.cs.byu.edu/fitnessfunction/
exhibit was held in the High Museum of Art and covered by NPR’s Studio 360\textsuperscript{2}. Again, the exhibit challenged visitor’s perceptions about art curation.

9.2 Utah County Art Gallery

DARCI’s interaction with the public didn’t end with \textit{Fitness Function}. In 2011, we submitted two digital images to the Utah County Art Gallery: Fall Photography and Digital Art Show. These images were selected from a pool of images created by DARCI using a blank white or black source image for the adjectives “happy”, “peaceful” or “scary”. For speed, the initial renderings DARCI created were for 256 x 256 resolution images. We expanded some of the renderings to 6000 x 4800, often changing them significantly, and made our selection from these expanded renderings. Thus, these submissions were more of a collaboration between us and DARCI. The two images we submitted were both created with the adjective “peaceful”, the first using a blank black source image (Figure 9.2(a)) and the second using a blank white source image(Figure 9.2(b)). The images were submitted under my name with no indication that a computer program was the primary artist. The image in Figure 9.2(a) won 2\textsuperscript{nd} place in the Digital Art category.

\footnote{http://www.studio360.org/story/175741-darci-computer-great-taste/}
9.3 Evolutionary Art, Design, and Creativity Competition

In 2013, we submitted several images to the GECCO Evolutionary Art, Design, and Creativity Competition held in Amsterdam. These images were the result of a collaboration with Derrall Heath to extend the capabilities of DARCI beyond rendering pre-existing images [42]. Heath’s expanded DARCI uses a semantic network to select nouns and adjectives that are associated with given concepts. The system then composes a collage of the selected nouns using a database of icons and a simple heuristic. Finally, the adjectives with the highest association are used to render the collage using the original rendering system. As with the Utah County Art Gallery, we selected the submitted images from a pool of candidates created by DARCI. Figure 9.3 shows the submitted images along with their source concept and a short description we provided for the competition to explain icon and adjective choices. The images were exhibited as finalists for the competition.
Figure 9.3: Images submitted to the Evolutionary Art, Design, and Creativity Competition held in Amsterdam as part of GECCO (Genetic and Evolutionary Computation Conference). The concepts used to create the pieces are listed under each image. (a) *War* incorporates concepts such as tank, gun, bomber and atom and renders them with a style that suggests explosive and bloody. (b) *Epic Drug Scandal* weaves a dizzy conception of icons such as pill, marijuana, medicine and syringe. (c) *Guilty Protest* combines concepts such as student, banner, crime and jail in a rendering designed to evoke a feeling of sadness. (d) *Murder* offers a dark, oppressive evocation of the grim reaper, electrocution and weapons. (e) *Artificial Intelligence* offers a quirky mix of conceptual proxies for intelligence, such as brain and school, with elements associated with artificial, like flower and lung, rendered to evoke the idea of light.

### 9.4 You Can’t Know My Mind

Later in 2013, we collaborated with Simon Colton as part of a festival celebrating computational creativity at the Galerie Oberkampf in Paris, France [14]. The festival, called *You Can’t*
Know My Mind\textsuperscript{3}, showcased several systems developed as part of research in computational creativity. Components of DARCI were incorporated into Colton’s *The Painting Fool* [13] to create a playful hybrid system capable of justifying its own artwork and reflecting upon it.

Visitors to the gallery had their photographs taken for reference. Then the hybrid system would determine its “mood” using sentiment analysis of current news highlights from *The Guardian*. The “mood” was, in turn, used to determine an adjective that the system would employ in producing a stylized portrait of the visitor. Then, assuming the system was in the proper “mood” to even paint, DARCI would select a non-photorealistic image filter and abstract background that best reflected the given adjective. The filter would then be applied to the subject on the background creating a “preconception” of the final portrait. Next *The Painting Fool* would digitally paint the “preconception” in a style dictated by its “mood”. DARCI would then analyze the resulting image to determine how well the final image reflected the desired adjective. Finally, using simple natural language templates, the hybrid system would return a description of the portrait including what it intended to produce, why it painted the way it did, and how it felt about its creation. Figure 9.4 shows an example of a portrait produced at the festival, with the accompanying reflection, that the system was pleased with. Figure 9.5 shows another example that the system was unhappy with.

\textsuperscript{3}http://www.thepaintingfool.com/galleries/you_cant_know_my_mind/
I was in a negative mood.
So I wanted to paint a bleary portrait.
I aimed to achieve something like this:

And this is my painting:

Overall, this is a very bleary portrait.
I guess my style has achieved roughly the bleary level I wanted.
I’m OK with that.
And I’m also pleased that the portrait is bleached, because that suits my mood.

Figure 9.4: A portrait produced by The Painting Fool in collaboration with DARCI at You Can’t Know My Mind in Paris, France. The accompanying text playfully describes the internal workings of the hybrid system. Image courtesy of Simon Colton.
I was in a reflective mood, and I had previously had trouble achieving a bright and warm portrait. So, I tried again with a different style. I aimed to achieve something like this:

And this is my painting:

Overall, this is quite a bright portrait. My style has significantly lowered the level of bright here. This is a miserable failure - I'm very unhappy about that.

Figure 9.5: *The Painting Fool*, in collaboration with DARCI, refusing to paint a portrait because of its mood. Image courtesy of Simon Colton.
Chapter 10

Conclusion

We have presented a new computational system, DARCI (Digital ARtist Communicating Intention), designed to explore computational creativity in the domain of visual art. Throughout this dissertation we have demonstrated three broad contributions of the system: 1) the contribution to the field of computational creativity in the form of an original system, new approaches to achieving autonomy in creative systems, and new practical assessment methods; 2) the contribution to the field of computer vision in the form of new image features for affective image annotation and a new dataset; and 3) the contribution to the domain of visual art in the form of mutually beneficial collaborations and participation in several art galleries and exhibits.

10.1 Contribution to Computational Creativity

In developing DARCI, we subscribe to several philosophies of creativity including Ritchie’s essential properties of creativity [77], Csíkszentmihályi’s systems model of creativity [21], and especially Colton’s creative tripod (including his emphasis on the perception of creativity) [12]. We have described how we adopt these philosophies into DARCI’s creative process while evaluating the system’s artifacts with a variety of assessment techniques. Thus, we grapple with both the weak and strong computational creativity viewpoints. In doing so, we present several approaches to improve system autonomy with minimal degradation to artifact quality. These approaches include using models of human opinion (image-adjective associations) as a fitness function in a genetic algorithm, using speciation in the genetic algorithm to improve
candidate artifact diversity, and appropriately combining multiple metrics when curating. We also present several generalizable survey approaches for obtaining human opinion in order to assess the creativity of a system from an audience’s standpoint. Ultimately, we document the progression of DARCI from a partially functional image annotator to an autonomous system capable of producing surprising artifacts competitive, in some ways, with those produced by human artists.

10.2 Contribution to Computer Vision

Current research in affective image annotation is still limited in both effectiveness and scope. Typically, research in this area is restricted to a small number of affective labels and rarely allows for multiple labels per image. This may be in part due to the lack of datasets with extensive affective labeling. Also, results in affective image annotation are usually not much better than random. Since DARCI operates on the notion that it can associate images with adjectives, most importantly affective ones, we are keenly interested in this topic of computer vision. In order to achieve satisfactory results for our research, we examine 197 global image features. Among these 197 image features are 64 new features that we introduce to the computer vision community. 52 of these new features are summary statistics of local interest point features prominent in computer vision. Using forward feature selection and the DARCI dataset, we select 51 features—half of which are our new features. Using a thorough comparison to state-of-the-art approaches across three established datasets, we show that our approach to affective image annotation is superior in most cases. We also introduce, justify, and publish a new dataset to the community that is derived from data we have collected to form DARCI’s image-adjective associations.

10.3 Contribution to the Visual Arts

Toward the end of this dissertation, we have described an unusual collaboration with Brigham Young Universities Visual Arts program. During this collaboration, we designed activities
to educate visual art students about DARCI’s creation process. In turn, many of these
students devised their own activities to incorporate their personal understanding of DARCI
into art works. Aside from creating some interesting pieces (see Chapter 8), many of these
students gained insights into their creation process that impacted their art going forward.
Furthermore, we have discussed several exhibitions in which DARCI and its artifacts have
impressed and sparked the curiosity of many.

10.4 Future Work

Despite the successes to date, DARCI is far from being complete with regards to its original
design goals. While limited in scope for this dissertation, DARCI is intended to eventually
produce wholly original images that convey any concept to their viewers. Ideally, it will be
able to determine the concepts and other sources of inspiration on its own, and even to justify
its creative process to others.

In collaborations with other researchers, we have already begun to see some progress
towards these goals. In particular, Heath is continuing to explore semantic models, inspired
by cutting edge natural language processing techniques, in order to generate more interesting
and semantically rich collages [41, 42]. His semantic models may also enable DARCI to more
accurately derive inspiration from media sources and more naturally articulate justification
for its art.

We also hope for more collaboration with other creative systems as we have done with
Colton’s The Painting Fool. It would be especially interesting to collaborate with systems
operating in other creative domains such as music and video games. For example, Johnson
has developed a system for discovering musical motifs from digital images [46]. It would be a
simple matter to have DARCI produce inspirational images for Johnson’s system. Also, Cook
is developing a system called ANGELINA that creates original video games [18]. DARCI
may be capable of producing artwork for backgrounds or asset textures within ANGELINA’s
games. Human artists commonly collaborate across artistic domains to great effect. Maybe the next step in computational creativity will be collaboration amongst digital artists.
References


