Autonomously Creating Quality Images

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

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.

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 (2007). Ritchie defines quality as the extent to which the artefact 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 artefacts 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 artefacts 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 artefacts. 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 artefacts.

Image Appreciation

It has been argued that the ability to appreciate and evaluate its own artefacts is necessary for a system to be considered creative (Colton 2008). 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 (Gevers and Smeulders 2000; Datta et al. 2006; Li and Chen 2009; Wang, Yu, and Jiang 2006; King ; Wang and He 2008), 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 (Fellbaum 1998) 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" (De Smedt, De Bleser, and

Nijs 2010). 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 (Norton, Heath, and Ventura 2010).

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 (Colton et al. 2010). 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 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:

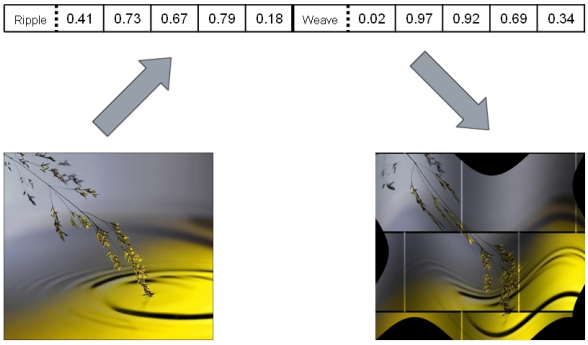


Figure 1: Sample genotype (list of image filters with parameters) and its effect on an image. "Ripple" and "Weave" are the names of two (of ninety-two) possible filters.

$$\text{Fitness}(g) = \lambda_A A(g) + \lambda_I I(g) \quad (1)$$

where g is an image artefact 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 artefact (here, G represents the set of all image artefacts). $\lambda_A + \lambda_I = 1$, 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:

$$A(g) = \sum_{w \in C} \alpha_w \text{net}_w(g) \quad (2)$$

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:

$$n = \sum_i \lceil 0.3 - |F_i^S - F_i^P| \rceil \quad (3)$$

Number of Sub-Populations	8
Size of Sub-Populations	15
Crossover Rate	0.4
Filter Mutation Rate	0.03
Parameter Mutation Rate	0.1
Migration Rate	0.2
Migration Frequency	0.1
Tournament Selection Rate	0.75
Initial Genotype Length	2 to 4 filters

Table 1: Parameters used for the evolutionary mechanism.

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:

$$I(g) = 1 - \begin{cases} \frac{\tau_d - n}{\tau_d} & \text{if } n < \tau_d \\ \frac{n - \tau_s}{|F| - \tau_s} & \text{if } n > \tau_s \\ 0 & \text{if } \tau_d \leq n \leq \tau_s \end{cases} \quad (4)$$

τ_d and τ_s are constants that correspond to the threshold for determining, respectively, when a phenotype is too different from or too similar to the source image. The values $\tau_d = 20$ and $\tau_s = 57$ were used here. $|F|$ is the total number of features analyzed, 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 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 change, Δf_i, for each parameter, i, is determined by one of the following two equations chosen randomly with equal probability:

$$\Delta f_i = (1 - f_i) \cdot rand\left(0, \frac{(|f| + 1) - |\Delta f|}{|f|}\right) \quad (5)$$

$$\Delta f_i = -f_i \cdot rand\left(0, \frac{(|f| + 1) - |\Delta f|}{|f|}\right) \quad (6)$$

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 $rand(x, y)$ 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).

Methods and Results

The evaluation of artefacts is very subjective, making an evaluation of DARCI non-trivial. Furthermore, the quality of the artefacts that DARCI produces can be judged based on two distinct criteria: how well the artefacts portray the synsets dictated by a commission, and how well the artefacts 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 artefact 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 artefacts. 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 artefacts, alongside several other artefacts, 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 2. We chose photographs in order to accentuate the impact of the non-photorealistic rendering tools available to DARCI. The

(a) Image A

(b) Image B

(c) Image C

Figure 2: The three source images used to evaluate the quality of DARCI's artefacts.

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 artefact (DARCI), one was our choice out of DARCI's top artefacts (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

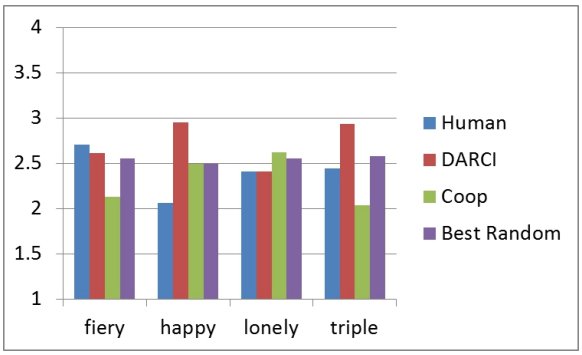


Figure 3: The average ranking for each synset across images A, B, and C for each of the four artefact sources: DARCI, Human, Coop, and Best Random. "triple" refers to the artefacts rendered with all three synsets. These results were obtained from 42 volunteers. Lower rank is better.

	Human	DARCI	Coop	Best Random
Average Rank	2.4067	2.7282	2.3194	2.5456

Table 2: The average ranking of the four artefact sources across all image-synset combinations. These results were obtained from 42 volunteers. Lower rank is better.

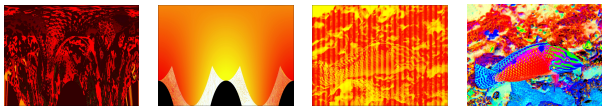
synset-image combinations can be found in Figures 4-7. In the online survey, volunteers were instructed to rank the four images for each synset-image combination according to how well they portrayed the synset(s) in question. In addition, we asked the volunteers to indicate which images they liked regardless 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 and Table 2. Figure 3 shows the average ranking for each synset across all three images for each of the four artefact sources just summarized: DARCI, Human, Coop, and Best Random (the lower the rank, the better). Table 2 shows the average ranking of each of the four artefact sources across all synsets and images. Table 3 shows which pairs of datapoints in the aforementioned figures are statistically significant—such pairs are denoted with an asterisk.

	fiery	happy	lonely	triple	all synsets
Human/DARCI	0.501	*	1.000	*	*
Human/Coop	*	*	0.155	*	0.217
Human/Best Random	0.286	*	0.326	0.339	0.0540
DARCI/Coop	*	*	0.132	*	*
DARCI/Best Random	0.691	*	0.298	*	*
Coop/Best Random	*	1.000	0.640	*	*

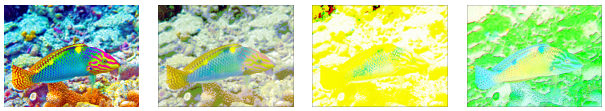
* p-value < 0.01

Table 3: Results of t-Test comparing all binary combinations of image sources for each synset. The "all synsets" column refers to Table 2. The other columns refer to Figure 3.



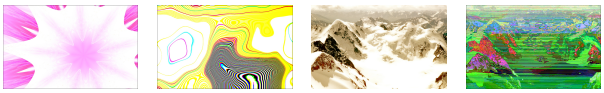
(a) DARCI (b) Best Random (c) Coop (d) Human

Figure 4: Image A rendered "fiery"—ranked left-to-right from most to least fiery.



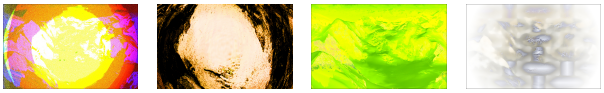
(a) Human (b) Coop (c) DARCI (d) Best Random

Figure 5: Image A rendered "happy"—ranked left-to-right from most to least happy.



(a) Best Random (b) Coop (c) Human (d) DARCI

Figure 6: Image B rendered "happy"—ranked left-to-right from most to least happy.



(a) Coop (b) Human (c) DARCI (d) Best Random

Figure 7: Image B rendered "fiery", "happy", and "lonely"—ranked left-to-right from most to least fiery, happy, and lonely.

Discussion

By looking at Table 2, we see that DARCI functioning autonomously does perform the worst of the artefact sources, but not dramatically so. Furthermore, Table 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 artefact source over another.

When looking at performance over individual synsets (Figure 3), we see that a more distinct preference is given to certain artefact sources over others. But, even in these cases, the source given preference varies from synset to synset. Looking at Figure 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 artefact 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 artefact 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 artefact source are shown in Figures 4-7.

While DARCI's solo artefacts often rank on par with human artefacts, the best random artefacts do as well. Furthermore, these partially random artefacts are sometimes ranked better than DARCI's. If these were totally randomly generated artefacts, 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 artefacts produced should be correlated with creativity (Ritchie 2007), and observing DARCI's top forty artefacts, it becomes clear that DARCI is accomplishing something better than random image generation. Figure 8 shows the 40 images DARCI chose to save while rendering image A as "fiery", while, for comparison, Figure 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 8 than in Figure 9.

Conclusions

If we assume that the human artists commissioned to produce artefacts 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 artefacts. 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 artefacts to facilitate visual communication with an audience. DARCI

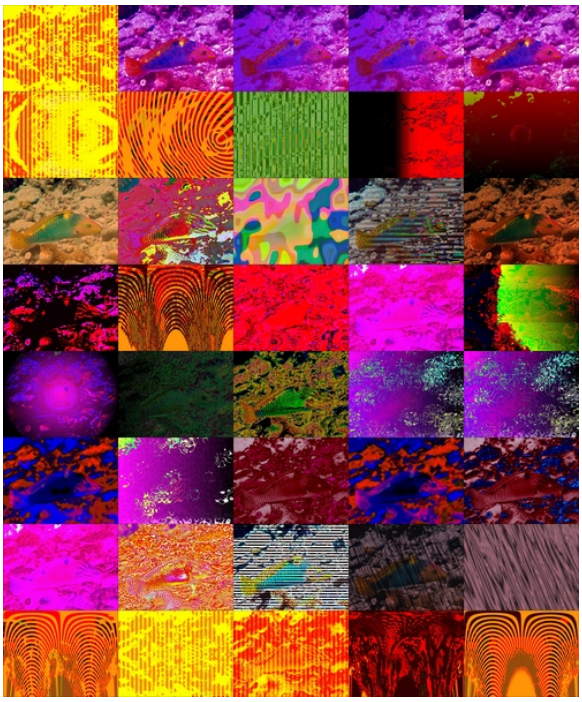


Figure 8: The top five "fiery" renderings for all eight subpopulations discovered by DARCI for image A (not ordered by fitness).

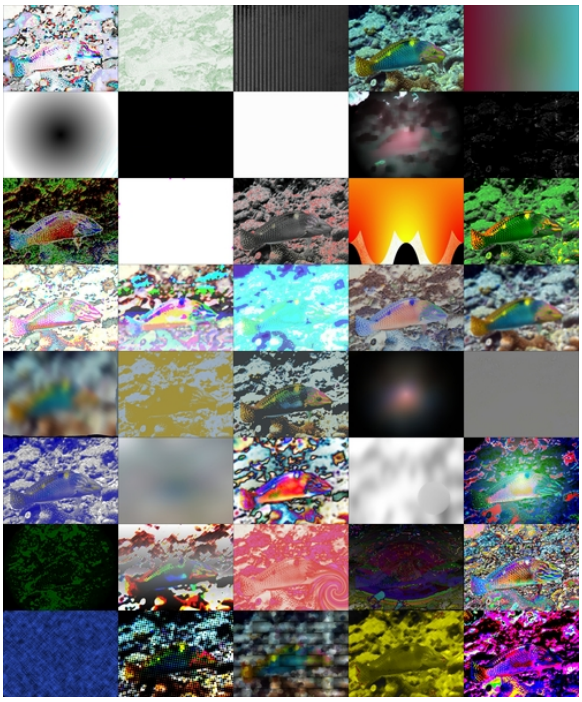


Figure 9: The forty randomly generated renderings of image A for the synset "fiery".

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 artefacts).

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.

References

Colton, S.; Gow, J.; Torres, P.; and Cairns, P. 2010. Experiments in objet trouvé browsing. *Proceedings of the International Conference on Computational Creativity* 238–247.

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

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

De Smedt, T.; De Bleser, F.; and Nijs, L. 2010. Nodebox 2. *Proceedings of the 16th International Symposium on Electronic Art (ISEA 2010)*.

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

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

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

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

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

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

Wang, W.-N., and He, Q. 2008. A survey on emotional semantic image retrieval. *Proceedings of the International Conference on Image Processing*.

Wang, W.-N.; Yu, Y.-L.; and Jiang, S.-M. 2006. Image retrieval by emotional semantics: A study of emotional space and feature extraction. *IEEE International Conference on Systems, Man, and Cybernetics* 4:3534–3539.