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

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

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, 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 rated at least as favorably as human counter parts across five aspects of creativity. Further, these artifacts tend to be more surprising and perceived as more difficult to produce than those created by human artists.

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* (al Rifaie and Bishop 2012)? 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, or to compare two different creative systems (Colton et al. 2014). The third problem is empirically grounding the ambiguous terminology that is commonly used to describe and assess creative systems (Brown 2014). The fourth problem is picking, or designing, the best methodology to actually carry out the assessment of a system (Jordanous 2014). The fifth problem, and one that is not addressed in detail by researchers in the field, is compensating for the effects of bias inevitably introduced by human evaluators when assessing creative systems.

While the researchers exploring these issues have presented tantalizing theoretical solutions, few have implemented practical solutions (a noted exception is Jordanous' meta-evaluation of existing evaluation methodologies (2014)). 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 DARCI system, attempting to address some of these problems using a specially designed, but generalizable, online human survey. Of note, we address the complication of bias introduced by human evaluators that is unaccounted for in current assessments of creativity.

There has been some reticence in the community towards conducting human surveys as a means of evaluation. Brown notes that human surveys often have wide variance making them difficult to incorporate into established models of creativity (2014). In a study comparing several methods of evaluation, Jordanous concludes that human surveys were the least correct of the methods she explored (2014). 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, ask participants to evaluate a variety of explicitly defined artifact qualities (that correspond to requirements for creativity) rather than asking them to directly evaluate the system's creativity.

Brown stresses the inadequacy of human surveys as empirically grounding assessments since we don't have an understanding of what the human responses mean (2014). In order to gain that understanding on some level, we develop a standard for judging the 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, in addition to the artifacts themselves, must be evaluated (2014). While our survey questions do focus on the artifacts, some are designed to glean opinions about DARCI's creative process. Unfortunately, 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 possible source of positive bias is potential viewer familiarity with computational creativity, or even DARCI itself, and a concomitant desire for 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. We then use this information to determine the effects of survey taker bias and adjust our conclusions from the survey accordingly.

DARCI and Artifact Creation

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 for rendering an image composition (produced by another subsystem) so that it will reflect a given description (selected from yet another subsystem).

DARCI 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 is different enough from the source image so as to satisfy the creativity requirement of originality while not being too different from the source image so as to satisfy the creativity requirement of functionality (Norton, Heath, and Ventura 2014).

To produce its artifact, DARCI first uses a system of genetic algorithms to build a pool of candidate artifacts from which to select the final rendering. Once these candidates have been created, DARCI uses a heuristic to rank them and then selects the top ranked candidate as the final artifact.

Candidate Artifact Creation

DARCI 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 image features with the adjective using standard backpropagation and a training set of hand-labeled images. 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. (2015). Many of these image features are the result of psychological studies analyzing the connection between color and various affective words (Ou et al. 2004; Wang, Yu, and Jiang 2006; Machajdik and Hanbury 2010). Others summarize local interest point data that is typically reserved for object detection in images (Norton, Heath, and Ventura 2015). Still other features come from a publicly available[1] set of widely accepted global image features (King, Ng, and Sia 2004).

[1] http://appsrv.cse.cuhk.edu.hk/~miplab/discovir/

Once the adjective ANN is trained, DARCI uses a genetic algorithm to discover the configuration and parameter settings of Photoshop-like filters for rendering the source image to reflect the given adjective. Candidate filter sequences are evaluated by applying them to the source image and 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 sub-populations.

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. To identify those artifacts that are not sufficiently unique, the system calculates the normalized cosine similarity between the 51-element feature vector of each potential candidate and the feature vector for each existing candidate. 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.

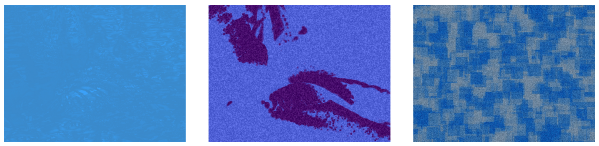
Once the candidate artifacts have been selected, another epoch of evolution is performed. This time a neural network we call the *novelty ANN* is trained to distinguish images novel to DARCI (the hand-labeled images mentioned previously) from those produced by the system (the pool of candidate artifacts). This process is similar to the process employed by Machado et al. in training NEvAr to create novel images (2007).

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. (2014). 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.

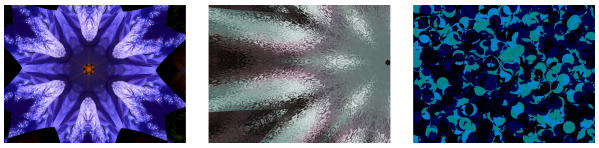
This process is repeated for several epochs, each adding increasingly varied images to the pool of candidate artifacts as the system attempts to optimize the changing fitness function. For our experiments, we perform a total of 8 epochs including the initial novelty-ANN-free 0^{th} epoch. Figure 1 illustrates how candidate artifacts vary from epoch to epoch during one experiment with the adjective "cold" using the image of Figure 2 as the source image.

Candidate Artifact Curation

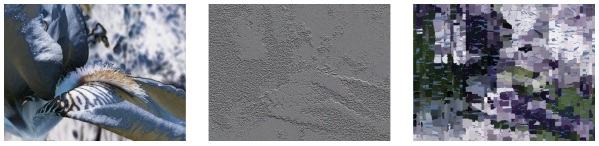
Once the candidate pool has been created, DARCI selects a single rendering to present as the finished product. Curating the candidates consists of two phases. In the first phase, DARCI ranks the candidates by their similarity to the source image and selects the top 10% (see Figure 3a - 3c), increas-



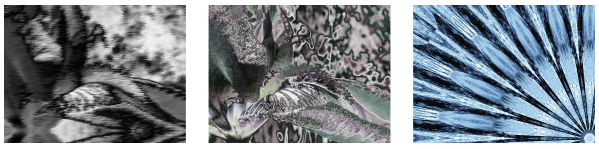
(a) Epoch 0



(b) Epoch 1



(c) Epoch 2

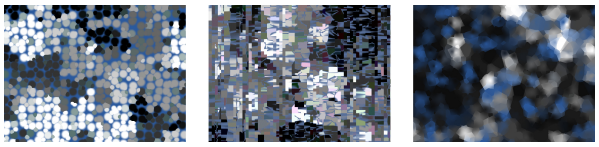


(d) Epoch 3

(e) Epoch 4

(f) Epoch 5

(g) Epoch 6



(h) Epoch 7

Figure 1: Sample artifacts from each epoch of the candidate building process for the adjective "cold" and source image in Figure 2. 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 2: The source image for all experiments in this paper.

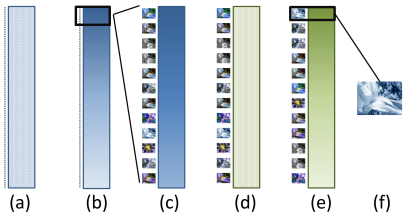
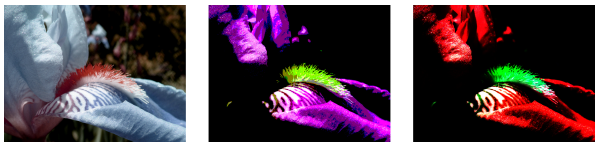


Figure 3: The curation process for selecting a final artifact from the pool of candidates (represented by a colored bar). (a) Each artifact in the pool of candidates is assigned a score of similarity to the source image (in this case Figure 2). (b) The candidates are ranked by this score (depicted by the bar's 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 how well they match the given adjective. (e) The artifacts are ranked by the new score. (f) The top image is selected as the final artifact to be returned.

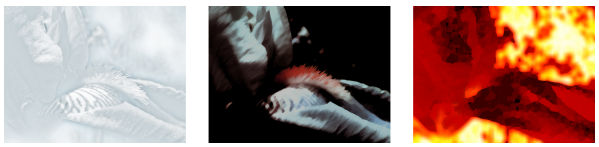
ing the chance that the final rendering will make noticeable use of the source image.

During curation, similarity to the source image is calculated to preserve the content, rather than the color, of the source image. Color usage has been shown to correlate with the affect of images (Wang, Yu, and Jiang 2006; Li and Chen 2009; Norton, Heath, and Ventura 2013), and we would actually like the color of the image to change in order to match the adjective description while keeping major objects within the source image recognizable. Therefore, similarity is calculated by first extracting a 1000-element histogram of visual words from the source image and each candidate artifact (visual words are quantized local image features commonly used in content-based image retrieval approaches (Sivic and Zisserman 2003)). The similarity between two images is calculated by taking the cosine similarity of the images' visual word histograms, as this similarity function has previously been used to successfully preserve the source image (Norton, Heath, and Ventura 2014).

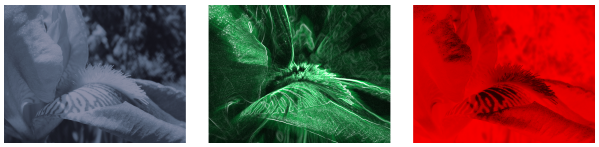
In the second phase of curation, DARCI ranks the remaining candidates by their association with the given adjective using the adjective ANN (see Figure 3d - 3f). The highest ranked image is then selected as the final artifact. This second phase occurs after over-filtered images have been removed in order to increase the chance that the final artifact reflects the given adjective and to reduce the possibility of returning an under-filtered image.



(a) Human 1



(b) Human 2

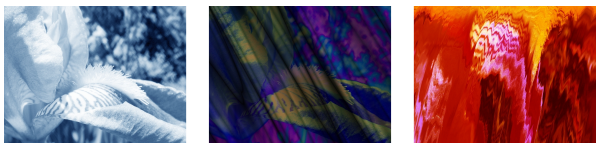


(c) Human 3

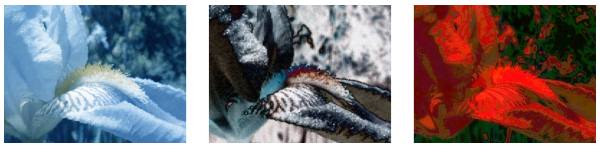
(d) Human 4

Figure 4: Renderings of Figure 2 created by four human artists. The renderings were created to depict, from left to right, the adjectives “cold”, “eerie”, and “violent”.

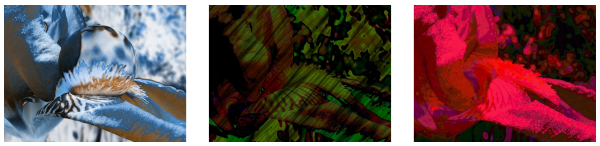
(a) *DARCI 1



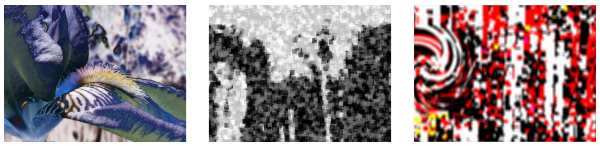
(b) *DARCI 2



(c) *DARCI 3



(d) *DARCI 4



(e) DARCI 5

(f) DARCI 6

(g) DARCI 7

Figure 5: Renderings of Figure 2 created by DARCI. The first four sets (*) were selected for the study. The renderings were created to depict, from left to right, the adjectives “cold”, “eerie”, and “violent”.

Commissions

For our experiments, we commissioned DARCI and four human volunteers to produce renderings of the photograph in Figure 2 that depict it as “cold”, “eerie”, and “violent”, respectively. These adjective were chosen because DARCI is able to associate them with images effectively (Norton, Heath, and Ventura 2015), 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 the human artists as similar as possible, human artists were restricted to a subset of tools found in software packages used for photo manipulation.

All four human volunteer artists have experience working with photo manipulation software, and, for grounding, they were shown examples of human-produced renderings from a previous study. The 12 images they produced for our study are shown in Figure 4.

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 increase output diversity across these commissions, the error threshold used in training the neural networks was varied for several commissions. To match the number of human commissions, we selected four of the artifacts DARCI produced for each adjective. We made the final decision to ensure varied artifacts and to eliminate potential outliers. Figure 5 shows all of the artifacts produced by DARCI and notes our chosen images.

Online Survey

In order to easily gather many responses, the survey was anonymous and online. To our knowledge, 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 Volunteers were presented with 15 pairs of images from which they would indicate their preferences. All images were created by applying random filters from DARCI's toolset to random source images and selecting intriguing and abstract creations from the thousands of random images. In order to limit the number of factors survey takers would be required to consider when making their selection, we paired images together that seemed similar in some respect. These image pairs were presented to volunteers in a random order with random labels. The labels indicated that one of the images was created by a human, and the other was created 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. Since the images were randomly labeled as "human" or "computer", unbiased volunteers should pick the "human" and "computer" options approximately equally.

Part 2 All volunteers were randomly assigned to one of three experiments: *blind*, *basic*, or *detailed*. The experiments were identical except for the amount of information that was presented to each volunteer. In all three experiments volunteers were given the following instructions:

> 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. The images were created using digital tools to modify a specific source photograph so that it reflected a given word.
>
> As an example, observe how an artist modified the following source photograph so that it reflected the word "happy".

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 which images were produced by DARCI. 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 produced its images. This description was followed by a simple one question quiz to assess comprehension.

Aside from the noted differences, the three experiments proceeded identically. Survey takers were presented with the source photograph (Figure 2), noted as such, and then six random images presented in random order: one image 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. To assist with these items, volunteers were always allowed to view the source photograph.

For all images, except the source, 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*)

For the source image we asked about all three adjectives, and omitted the three items that referred to the source.

Participants were not asked to explicitly assess the creativity of artifacts since personal notions of creativity vary widely. Instead, these five items were chosen to succinctly capture certain qualities required to attribute creativity to a system via the artifacts it produces, and to a small extent, its creative process. Norton et al. have shown that a similar set of Likert items are reliable (using Cronbach's alpha) and correlate with participants' opinions of creativity as measured by an additional Likert item explicitly for "creativity" (2013).

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*) (2008), Ritchie's 18 criteria defined by functions of *quality*, *novelty*, and *typicality* (2007), Jordanous' 14 components of creativity (2012), 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.

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. 100 volunteers were assigned to the *blind* experiments, 111 to the *basic* experiment, and 106 to the *detailed* experiment. For 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.

Bias

A volunteer's bias was calculated by subtracting the number of images they preferred labeled with "computer" from those labeled with "human" in the first part of the survey. Thus, a positive score indicates a bias in favor of humans. 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 was 0.901 with a standard error of 0.185, indicating a small but substantial bias either towards humans or against DARCI.

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. These results, with standard error, can be seen in Figure 6.

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 *blind*, *basic*, and *detailed* experiments. A positive correlation between bias and a particular item for a given artist (human or DARCI) would indicate that a bias towards humans (or against DARCI) is correlated with an increase in the item score for the artist. Table 1 shows these correlation values and their p-values (calculated with a two tailed Student's t-test) for the three experiments.

Only the *detailed* experiment contained a correlation that was statistically significant to $p < 0.05$. That was a positive correlation with the *difficult* item in human produced images. This means that volunteers with a bias towards humans tended to give humans a boosted score for *difficulty* when they understood how DARCI produced images. Even though none of the other correlations were statistically significant, it should be noted that in the two most informed experiments, the correlations were generally more positive towards humans and more negative towards DARCI (as one might expect). But, the lack of significance indicates that bias did not have a substantial impact on most results.

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 1). 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 in favor of humans tended to give all images a lower score when they didn't know who produced them. Perhaps these volunteers were concerned that an image might be produced by DARCI. It would be interesting to investigate this phenomenon in future studies.

Human	*blind*		*basic*		*detailed*	
Item	r	p-value	r	p-value	r	p-value
like	-0.087	0.399	0.044	0.591	0.160	0.357
adjective	-0.089	0.389	-0.011	0.892	-0.059	0.737
surprising	-0.043	0.677	0.123	0.130	0.179	0.303
difficult	0.019	0.851	0.102	0.208	0.428	0.010
use	-0.154	0.133	0.050	0.539	0.295	0.085

DARCI	blind		basic		detailed	
Item	r	p-value	r	p-value	r	p-value
like	-0.047	0.651	-0.060	0.465	0.070	0.690
adjective	-0.076	0.462	-0.0117	0.150	-0.012	0.944
surprising	-0.045	0.661	0.068	0.405	-0.212	0.222
difficult	-0.098	0.342	-0.036	0.662	-0.117	0.502
use	-0.073	0.480	0.002	0.983	0.029	0.869

Table 1: The Pearson correlation coefficient, r, and associated p-value, between volunteer bias and item scores for the three experiments (*blind*, *basic*, *detailed*). Positive correlation indicates that a bias towards humans is correlated with an increase in item score.

Evaluation

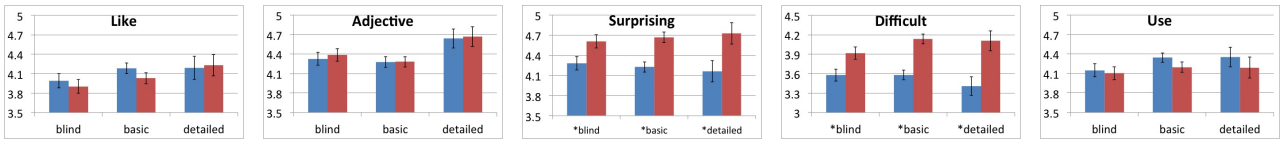
The average scores of the source image across its four Likert items were 5.873 (*like*), 2.260 (*cold*), 1.870 (*eerie*), and 1.377 (*violent*). Looking at Figure 6b we see that both humans and DARCI were able to reflect the adjectives more effectively in their artifacts than did the original source (though at the cost of a lower "like" score).

While the Likert scale is one of the most common evaluation tools used in psychology and marketing research, it has come under criticism for the unintended effects that it can introduce, including cultural biases, memory effects, and the loss of individual subjectivity when the scale is averaged across participants. Recently, it has been demonstrated that ranking or preference questionnaires have fewer negative effects (Yannakakis and Hallam 2011) and that converting from a rating scale to preferences can reduce some of the undesired effects of Likert questionnaires (Martínez, Yannakakis, and Hallam 2014).

To augment the rating-based results of Figure 6, individual survey takers' preferences were calculated from their Likert scores. For each Likert item and for each participant, we performed a pairwise comparison of all images reviewed by the survey taker. We tabulated which images scored higher (were preferred) and when ties occurred in these pairwise comparisons. To summarize the results, we have indicated the percentage of all pairwise tests for each item where human art was preferred, DARCI's art was preferred, and when ties occurred (Figure 7).

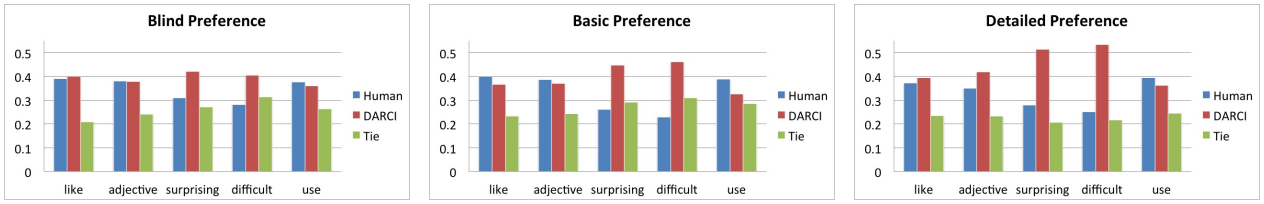
Looking at Figures 6 and 7 we see that DARCI clearly scored higher than human artists in the *surprising* and *difficult* categories while humans did not score substantially higher than DARCI in any category. These trends persisted across all experiments despite the overall human bias of the volunteers. In Figure 6, statistically significant differences ($p < 0.05$ using a two tailed Student's t-test) between humans and DARCI are starred (*).

While purely quantitative, these results suggest that



(a) (b) (c) (d) (e)

Figure 6: The average scores of each Likert item across all artifacts produced by either humans (blue, left) or DARCI (red) for each group of experiments (with standard error). (*) indicates statistical significance between human and DARCI results.



(a) (b) (c)

Figure 7: The result of every pairwise test, after converting Likert ratings to pairwise preferences. For each survey taker, pairwise tests were conducted between every combination of images produced by a human and those produced by DARCI.

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 given details about DARCI's creation process responded to its artifacts very similarly to how those volunteers not given the details responded—understanding how DARCI functioned did not diminish the way the artifacts were perceived.

Somewhat surprisingly, the additional information provided to some of the survey participants had minimal effect on their responses. There was no statistically significant difference between the results of any of the experiments except between the *basic* and *detailed* experiments in the *adjective* category (note the increased scores for both DARCI and human for the *detailed* experiment of Figure 6b). In this one case, understanding how DARCI produced artifacts influenced how volunteers perceived the meaning of the images produced by both DARCI and humans. Since the *detailed* group was told that DARCI learned to associate images with words through training by human teachers, volunteers may have realized that all of the images they were evaluating were essentially examples of what their peers associated with the given adjective. In other words, we suggest that volunteers were incorporating Jordanous' 'social interaction and communication' component into their evaluation.

Table 2 shows the top six images in each category for the three experiments. Refer to Figures 4 and 5 to view the actual images. Of note, DARCI's artifacts have a slightly greater representation amongst the highly rated images.

Conclusions

We have described recent improvements to a computational system, DARCI, that generates renderings of images so that they reflect an adjective and have presented a human-survey-based instrument designed to evaluate DARCI's artifacts and creation process while taking participant bias into consideration. The instrument 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 in question.

By analyzing the survey results, we have shown that across each of our criteria for creativity, DARCI's artifacts were rated comparably to artifacts produced by humans. Of note, DARCI's images were generally considered more surprising and more difficult to create than their human counterparts. DARCI's performance in the evaluation persisted even when volunteers (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. We must also acknowledge that the artifacts were very specific in nature and the human artists were heavily restricted in there creative process 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 baseline for testing future improvements to the system.

blind	*basic*	*detailed*
Like		
DARCI 4 “cold”	DARCI 4 “cold”	DARCI 3 “cold”
DARCI 3 “cold”	Human 1 “cold”	Human 3 “eerie”
Human 1 “cold”	Human 4 “violent”	Human 1 “cold”
Human 2 “cold”	DARCI 3 “cold”	Human 4 “violent”
Human 4 “cold”	Human 4 “eerie”	Human 2 “cold”
DARCI 2 “eerie”	Human 4 “cold”	DARCI 2 “eerie”
Adjective		
DARCI 2 “cold”	DARCI 1 “cold”	DARCI 3 “cold”
DARCI 4 “cold”	DARCI 3 “cold”	Human 2 “cold”
DARCI 3 “cold”	DARCI 2 “cold”	DARCI 2 “cold”
DARCI 1 “cold”	Human 2 “eerie”	Human 3 “eerie”
Human 2 “violent”	Human 2 “cold”	DARCI 1 “cold”
Human 2 “cold”	DARCI 4 “cold”	DARCI 4 “eerie”
Surprising		
Human 3 “eerie”	DARCI 2 “violent”	DARCI 1 “violent”
DARCI 2 “violent”	DARCI 1 “violent”	Human 3 “eerie”
DARCI 1 “violent”	Human 3 “eerie”	DARCI 2 “violent”
DARCI 2 “eerie”	DARCI 2 “eerie”	DARCI 1 “cold”
DARCI 4 “cold”	DARCI 4 “cold”	Human 2 “violent”
Human 2 “violent”	Human 2 “violent”	DARCI 4 “eerie”
Difficult		
DARCI 1 “violent”	DARCI 2 “violent”	DARCI 1 “violent”
DARCI 2 “violent”	DARCI 1 “violent”	Human 3 “eerie”
DARCI 2 “eerie”	Human 3 “eerie”	DARCI 2 “violent”
Human 2 “violent”	DARCI 2 “eerie”	Human 2 “violent”
Human 3 “eerie”	Human 2 “violent”	DARCI 2 “eerie ”
Human 2 “eerie”	DARCI 4 “cold”	DARCI 4 “violent”
Use		
Human 3 “eerie”	DARCI 4 “cold”	Human 4 “eerie”
Human 4 “eerie”	Human 1 “cold”	Human 1 “cold”
DARCI 4 “cold”	Human 4 “violent”	Human 3 “eerie”
DARCI 3 “cold”	Human 4 “cold”	DARCI 1 “cold”
Human 1 “cold”	DARCI 1 “cold”	DARCI 3 “cold”
Human 4 “violent”	DARCI 2 “eerie”	DARCI 1 “eerie”

Table 2: The top six images (based on Likert rating) for each item across the three experiments. Refer to Figures 4 and 5 to view images.

References

al Rifaie, M., and Bishop, M. 2012. Weak vs. strong computational creativity, computing, philosophy and the question of bio-machine hybrids. In *Proceedings of the AISB Symposium on Computing and Philosophy*.

Brown, O. 2014. Empirically grounding the evaluation of creative systems: Incorporating interaction design. In *Proceedings of the International Conference on Computational Creativity*.

Colton, S.; Pease, A.; Corneli, J.; Cook, M.; and Llano, T. 2014. Assessing progress in building autonomously creative systems. In *Proceedings of the International Conference on Computational Creativity*.

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

Jordanous, A. 2012. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation* 4:246–279.

Jordanous, A. 2014. Stepping back to progress forwards: Setting standards for meta-evaluation of computational creativity. In *Proceedings of the 5th International Conference on Computational Creativity*.

King, I.; Ng, C. H.; and Sia, K. C. 2004. Distributed content-based visual information retrieval system on peer-to-peer network. *ACM Transactions on Information Systems* 22(3):477–501.

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

Machado, P.; Romero, J.; and Manaris, B. 2007. Experiments in computational aesthetics: An iterative approach to stylistic change in evolutionary art. In Romero, J., and Machado, P., eds., *The Art of Artificial Evolution: A Handbook on Evolutionary Art and Music*. Berlin: Springer. 381–415.

Machajdik, J., and Hanbury, A. 2010. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the International Conference on Multimedia*, 83–92.

Martínez, H. P.; Yannakakis, G. N.; and Hallam, J. 2014. Dont classify ratings of affect; rank them! *IEEE Transactions on Affective Computing* 5:314–326.

Norton, D.; Heath, D.; and Ventura, D. 2013. Finding creativity in an artificial artist. *Journal of Creative Behavior* 47(2):106–124.

Norton, D.; Heath, D.; and Ventura, D. 2014. Autonomously managing competing objectives to improve the creation and curation of artifacts. In *Proceedings of the 5th International Conference on Computational Creativity*.

Norton, D.; Heath, D.; and Ventura, D. 2015. Annotating images with emotional adjectives using features that summarize local interest points. *IEEE Transactions on Affective Computing, in submission*.

Ou, L.-C.; Luo, M. R.; Woodcock, A.; and Wright, A. 2004. A study of colour emotion and colour preference. Part I: Colour emotions for single colours. *Color Research and Application* 29:232–240.

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

Sivic, J., and Zisserman, A. 2003. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*, volume 2, 1470–1477.

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

Yannakakis, G. N., and Hallam, J. 2011. Rating vs. preference: A comparative study of self-reporting. *Affective Computing and Intelligent Interaction* 6974:437–446.