To Appear In Proceedings of the 2nd International Conference on the Theory and Practice of Natural Computation, pp. 18 pages 2013. Springer.

Boosting Interactive Evolution using Human Computation Markets

Joel Lehman and Risto Miikkulainen

The University of Texas at Austin, Department of Computer Science, Austin, Texas
{joel,risto}@cs.utexas.edu

Abstract. Interactive evolution, i.e. leveraging human input for selection in an evolutionary algorithm, is effective when an appropriate fitness function is hard to quantify yet solution quality is easily recognizable by humans. However, single-user applications of interactive evolution are limited by *user fatigue*: Humans become bored with monotonous evaluations. This paper explores the potential for bypassing such fatigue by directly purchasing human input from human computation markets. Experiments evolving aesthetic images show that purchased human input can be leveraged more economically when evolution is first seeded by optimizing a purely-computational aesthetic measure. Further experiments in the same domain validate a system feature, demonstrating how human computation can help guide interactive evolution system design. Finally, experiments in an image composition domain show the approach's potential to make interactive evolution scalable even in tasks that are not inherently enjoyable. The conclusion is that human computation markets make it possible to apply a powerful form of selection pressure mechanically in evolutionary algorithms.

1 Introduction

A critical component of any evolutionary computation (EC) experiment is selection, i.e. how the parents of the next generation are chosen from the current population. In particular, the success of a particular EA in a given domain often depends upon choosing an appropriate *fitness function* to guide search. That is, for an EA to produce a solution, the fitness function that is optimized must induce a sufficiently smooth gradient of increasing fitness that leads from the random individuals in the initial population to a solution. However, intuitive choices for fitness functions may often fail to identify the intermediate steps that lead to the solution [10, 4], and some concepts intuitive to humans remain difficult to quantify algorithmically [17, 19].

For example, creating an algorithmic characterization of aesthetic appeal to automate evolving aesthetic artifacts is a compelling [11, 3, 13] yet unfulfilled endeavor [17, 13]. In such cases, one way to bypass this lack of an algorithmic measure is through interactive evolutionary computation (IEC; [19]), wherein humans act as a fitness function, actively selecting which solutions to evolve

further. The insight is that humans may be able to evaluate a characteristic even when it cannot be mechanically recognized or rigorously defined.

However, a significant problem in IEC is user fatigue: A single user can only perform so many evaluations before becoming tired or bored [19]. A recent solution to this problem is to create collaborative IEC websites whereby without financial incentive users cooperate to evolve complex artifacts they could not have evolved alone [17, 2, 12]. The idea is that although a single user may become fatigued, if that user *publishes* their work on the website, other users can choose to further *extend* that published work. Over time, the artifacts that are published can accumulate and form a branching phylogeny of diverse and interesting content [17].

This approach is economical and promising when task domains are inherently enjoyable, e.g. creative domains like open-ended image, shape, or music evolution [17, 2, 12]. However, when attempting to apply the approach to arbitrary domains there are two significant limitations: (1) sustained evolution for many generations depends upon the task domain being engaging enough to continually draw a sufficient volume of volunteer users, and (2) implementing the idea requires creating the non-trivial system architecture that composes a collaborative evolution website, e.g. architecture that supports creating and handling user accounts, facilitating discovering and rating artifacts, and evolving and publishing artifacts.

An interesting potential solution to these problems is provided by human computation markets (HCMs). In these markets it is possible to pay for human input in arbitrary tasks and thereby keep humans motivated even when the task is not particularly rewarding itself. This paper explores whether HCMs can be effectively used for this purpose, through an approach called HCM+IEC that uses HCMs to perform selection in an interactive evolutionary algorithm.

The paper focuses on three ideas: First, even if the domain to be used with IEC is itself enjoyable and engaging (e.g. evolving aesthetic images), IEC websites face the bootstrapping problem common to all user-generated content sites. That is, at such a site's launch, when attracting users is most important, the site is *least* engaging due to lack of content. Thus the first contribution of this paper is to suggest that markets for human computation can help overcome this bootstrap problem: Initially users can be paid to generate content. For this reason, experiments with such an aim apply IEC+HCM in an image evolution domain representative of those often explored by collaborative IEC websites. The results show that human computation can be more efficiently leveraged if a computational aesthetic measure [11] first algorithmically generates an interesting diversity of images upon which humans can further elaborate.

Second, when designing an IEC website or a single-user IEC system, often many design decisions about the underlying algorithm must be made that will significantly impact the quality of the system's output. Problematically however, such important decisions often are guided only by the preferences and intuitions of the system designers. The second contribution of this paper is thus to suggest that IEC+HCM can be applied to conduct controlled experiments that measure

the impact of a design decision on the quality of an IEC system's products. Experiments in the same image evolution domain show that removing a significant feature results in measurably less aesthetically pleasing pictures, thereby demonstrating the potential for IEC+HCM to facilitate principled IEC system design.

Third, there are EC problems that could benefit from large-scale human selection but for which a collaborative IEC website will not be a feasible solution. That is, most current IEC websites rely on self-directed users to produce content, and such content is produced irregularly and only to the extent that volunteer users *enjoy* evaluating artifacts in the domain. Thus the third contribution of this paper is to demonstrate how IEC+HCM can be used instead of a collaborative IEC website in one such condition, i.e. when the task domain is not enjoyable.

The conclusion is that HCMs offer a mechanism for converting money into a powerful form of selection pressure that may prove a valuable tool for interactive evolution.

2 Background

In this section, the foundational technologies applied in the experiments in this paper, including interactive evolution, human computation, and heuristics for evolving impressive artifacts, are reviewed.

2.1 Interactive Evolution

Applying human judgment to perform selection in an evolutionary algorithm is called interactive evolutionary computation (IEC; [19]) and is motivated by the difficulty in quantifying intuitive concepts that are readily recognized by humans (e.g. aesthetic appeal), and also by the impressive examples of human-directed breeding (e.g. the wide variety of domesticated dogs or the increased potency of human-bred agricultural crops). While IEC has been explored in the context of single-user applications [19], collaborative websites [17, 2], and online video games [6, 16], it has only been superficially explored in the context of HCMs [1], i.e. websites that facilitate paying human users to complete small tasks that cannot be easily algorithmically automated.

Note that although both combine human intution with evolutionary algorithms, IEC is distinct from Human-Based Genetic Algorithms (HBGAs; [9]). In HBGAs, humans perform not only selection (as in IEC), but implement *all* genetic operators and additionally serve as the substrate for genetic representation. That is, a human participating in a HBGA might recombine two existing pictures by *drawing* a new image that combines high-level features of both, instead of *breeding* together pictures generated by an underlying computational genetic system as in IEC. While HBGAs have been previously combined with HCMs [24, 23], they impose the requirement that humans be able to construct and manipulate the artifacts being evolved. In other words, often recognizing a promising artifact is easier than creating one. Thus, a potential advantage of

IEC is that one can breed complex artifacts (e.g. neural networks or complex pictures) without understanding their construction. In this way, IEC can enable humans without expert knowledge to aid in solving difficult computational problems.

Supporting this idea, previous studies with IEC have demonstrated its promise for evolving complex structures, e.g. significantly increasing efficiency when evolving artificial neural networks (ANNs) that control mobile robots [5, 22]. Also, in situations where large numbers of evaluations are impractical, it has been shown that IEC can make problems more tractable [5].

A representative example of a scalable IEC approach is given by the Picbreeder website [17], which encourages indirect user collaboration to evolve aesthetic images. On the site, users can discover, rate, and extend (through further interactive evolution) previously evolved images that are represented by compositional pattern producing networks (CPPNs; [18]). Note that CPPNs are feed-forward ANNs with an extended set of activation functions chosen for the regularities they induce [18]. The success of Picbreeder in evolving a wide variety of interesting and complex images likely stems from combining together an open-ended genetic encoding, an open-ended domain, and a powerful form of selection pressure (i.e. human judgment). Thus, similarly designed websites may be one path to large-scale IEC and compelling evolved artifacts.

However, evolution in such websites is typically undirected (i.e. driven by users' whims on what to create) and public (i.e. driven by users' ability to discover and elaborate upon existing content); for commercial IEC applications the ability to more directly guide the evolutionary process may be important and additionally it may be necessary for evolved content to be kept private (i.e. not stored such that all content is publicly accessible).

These limitations motivate exploring new approaches for large-scale IEC. A promising resource that can be leveraged for such purposes is human computation, which is reviewed next.

2.2 Human Computation

While the range of tasks solvable by computers continues to expand, there remain tasks that are challenging to solve computationally but are trivial for humans to solve. Examples of such tasks are recognizing written text [21], identifying objects in images [20], or evaluating aesthetic appeal [13]. This asymmetry motivates leveraging *human computation* [21] to automatically integrate human insight into algorithmic processes. Such human computation can often be made more scalable by employing *crowdsourcing* [8, 15], whereby many small contributions from a diffuse group of people (often online) are aggregated.

For example, while CAPTCHAs (Completely Automated Public Turing test to tell Computers and Humans Apart) separate humans from machines by generating tasks that are easily solvable by humans but difficult for machines, the widely-deployed reCAPTCHA system [21] acts as a CAPTCHA while at the same time leveraging human computation to transcribe words from old books. That is, part of each word identification task posed by reCAPTCHA to its users

is not machine generated, but is an image of a phrase that algorithmic character recognition struggled to classify automatically. Similarly, "games with a purpose" (GWAP; [20]) are designed such that human enjoyment results from deriving and verifying solutions to problems that are not yet solvable computationally; in this way, game players cooperate to create tagged data sets as a byproduct of an enjoyable experience [20].

ReCAPTCHA and GWAPs show that sometimes users can be enticed to generate useful computation without economic incentive. However, it is unclear how to transform an arbitrary human computation task into an enjoyable or necessary process such that the task's solution is a byproduct. Additionally, rather than wrap a task in a cleverly designed game and attempt to attract volunteers to play it, it may be simpler or cheaper sometimes to simply *pay* a human to perform the desired task through a HCM. The most well-known such marketplace is the Amazon Mechanical Turk (AMT; [7]), which is the system used in the experiments in this paper.

AMT exposes an interface to programmers that allows them to upload human intelligence tasks (HITs) which specify a desired task, an interface for humans to perform it, and the monetary reward for successfully completing the task. Once a human completes the HIT, the results can be queried and approved so that the human user can be paid. In this way, markets for human computation like AMT allow seamless integration of algorithms with arbitrary human input through economic exchange.

The next section reviews an algorithmic aesthetic measure that seeds evolution in some experiments in this paper.

2.3 Evolving Impressive Artifacts

One dimension of what humans appreciate as impressive is the perceived amount of design effort necessary to create an artifact. In other words, it is *easy* to recognize how *difficult* an impressive artifact was to create [11]. For example, intuitively it is easier to recognize a good novel than it is to write one, and it is easier to perceive a back-flip than it is to perform one. Supporting such an idea, artifacts often appear less impressive if they require significant effort to create but such effort is not readily apparent, e.g. a painting of something trivial that is entirely indistinguishable from a photograph. Conversely, when tasks are trivial, the disparity in effort between recognition and creation is much less; for example, reading a novel composed of random words may take more effort than actually writing one. Put another way, it is easy to verify the difficulty in creating an impressive artifact. Interestingly, characterizing impressiveness this way parallels the idea of NP-completeness: Solutions to NP-complete problems are easily verified but difficult to derive.

Although a philosophical description of impressiveness may be thought-provoking, its application to computational experiments is limited unless it can be quantified. Lehman and Stanley [11] introduced two heuristics for estimating algorithmically the design effort necessary to recreate an artifact: rarity and recreation effort. That is, because the ability to perform a back-flip is rare, such rarity may

indicate that back-flips are an impressive act. Similarly, the relative rarity of a noticeable property or combination of such properties in a wider space may hint that they are impressive; for example, an image with a symmetric tessellating pattern may be rare and thus potentially impressive. A more rigorous metric, although more computationally expensive, is recreation effort, i.e. the amount of effort required to design a similar artifact from scratch. The rarity heuristic, which is computationally easier to apply, was shown to largely agree with the recreation effort heuristic, and is therefore used in this paper [11].

While such algorithmic heuristics may not always agree with human intuition about how impressive an artifact is, they may still provide an automatic means to generate an interesting diversity of artifacts. This observation motivates applying the impressiveness metrics in the experiments in this paper in hopes of seeding search to more economically leverage human input. That is, in some experiments users from AMT are presented purely computationally-evolved impressive artifacts in the first generation. In particular, such experiments in this paper explore the same image evolution domain as Lehman and Stanley [11] although impressiveness metrics could be adapted to other domains, i.e. the general concepts are not particular to evolving images.

Fig. 1: **Comparing random and impressive images.** The images shown in (a),(b), and (c) are representative of images generated by random genomes in the image evolution domain, while the images shown in (d),(e), and (f) are examples of images evolved through the impressiveness metrics. Importantly, the impressive images differ qualitatively and are noticeably more complex than the random images.

To facilitate measuring impressiveness in the image evolution domain, Lehman and Stanley [11] compiled a list of features relevant to human vision, such as the level of symmetry in an image, an image's brightness, and how compressible an image was by different compression algorithms. Next, the rarity of different settings of simple combinations of these features was measured by sampling a space of images [17]. Finally, a search process driven to explore combinations of features was run, and its products screened by the rarity heuristic to cull the most impressive. The setup is described more exhaustively by Lehman and Stanley [11]. Note that other automated content-generation methods could have been substituted; what is most important for the purpose of this paper is that the impressiveness metrics enable automatic evolution of images that are more interesting than the random genomes that would otherwise seed evolution. For

comparison, examples of impressive images and those generated from random genomes are shown in figure 1.

3 Approach

While previous approaches to IEC are limited by user fatigue or require that a domain be enjoyable to attract users, the approach in this paper, IEC+HCM, avoids such issues by paying users for performing IEC evaluations through a HCM (figure 2). Of course, the trade-off is that with IEC+HCM there is an explicit economic cost for each evaluation.

The particular HCM applied here is AMT. Recall that AMT provides a computational interface for posting small computational tasks with a set monetary reward. Importantly, the AMT interface can be applied to automate IEC tasks (e.g. by presenting a user with an artifact or behavior and querying for evaluation via a web interface). Thus, tasks can be mechanically created and uploaded to AMT, results collected, and participants paid, without additional human oversight. In this way the methodology can potentially scale to arbitrary limits given enough money and available users in the HCM, thereby overcoming previous limitations to easily implementing IEC in any domain on a large scale. This paper evaluates the feasibility of the IEC+HCM approach: The experiments included 726 unique AMT users completing $2,300$ HIT evaluations. If successful, it should be eventually possible to scale the approach by two to three orders of magnitude.

Importantly, there are many potential ways to combine a HCM with an IEC algorithm. One design decision is how tasks should be divided. For example, a task sent to a HCM for completion could consist of evaluating only a single artifact, evaluating the evolutionary algorithm's entire population, or guiding multiple generations of IEC evolution (i.e. evaluating multiple populations in sequence). In this paper, tasks were divided into evaluations of a single population each, similarly to how a user influences a single generation of evolution in most single-user IEC applications [19].

Another important decision is what type of input should be gathered from human users; such input could consist of only which artifact in a population was most preferred, or could require individually rating each artifact. Individual ratings were gathered in this paper to enable comparisons between generations and runs, and to encourage greater deliberation during evaluation.

A final aspect of combining IEC with a HCM is how user evaluations of artifacts guide evolution. For example, if each user independently guides evolution for many generations, their most-preferred artifacts could seed the initial population of future tasks, similarly to how sites such as Picbreeder work [17]. However, in the approach in this paper each user evaluates only one generation at a time, and multiple independent evaluations of the same population are combined together to allocate offspring for the next generation. To avoid averaging out individual preferences, children are allocated to artifacts in proportion to how many users rate them most-highly (instead of simply averaging each artifact's ratings).

Thus while other approaches to IEC+HCM may also be viable, the approach described here reasonably combines IEC with HCMs, and its design decisions form a coherent methodology.

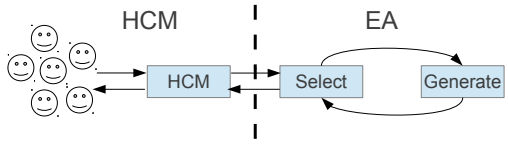

Fig. 2: **The IEC+HCM approach.** During each generation of evolution, when evaluating the population the EA uploads evaluation tasks to the HCM to be completed by human users. The results guide which parents reproduce to form the next generation. In this way, an EA can be driven by the human judgment of many non-experts.

4 Experiments

In following sections, experiments are presented that apply IEC+HCM to two domains. Results in evolving aesthetic images are first presented as a domain characteristic of collaborative IEC websites driven by user volunteers [17]. The second domain evolves only compositions of image layouts (which removes the potential for engaging novelty) as an exemplar for where economic incentives are crucial for success.

4.1 General Experimental Setup

For all experiments, human computation was purchased through AMT and a set price of $0.05 USD was paid per user completion of a task. A standard genetic algorithm was applied with a small population size (nine individuals) characteristic of many IEC domains [19]. All runs consisted of ten generations, with one task uploaded per generation, and ten independent runs of each compared method were performed to enable statistical comparisons.

Tasks uploaded to AMT contained nine images, i.e. the entire population, and required users to rate each image's aesthetic appeal on an integer scale from one to five (where five is the best). Because aesthetic judgment is subjective and varies between individuals, each task was evaluated by five separate AMT users to get a more representative sample. In particular, individuals were selected proportionally to how many users (from the five total) rated them most highly among the nine presented images (i.e. only a user's highest-rated images would contribute to selection). Preliminary experiments demonstrated that simply averaging the ratings was less effective than this approach because many

interesting artifacts were divisive, i.e. they were rated high by some and low by others, and averaging thereby would have resulted in selection for mediocre images that received lukewarm ratings from most users.

4.2 Evolving Aesthetic Images with IEC+HCM

The first two experiments explored an image evolution domain that implements an encoding similar to the Picbreeder collaborative IEC website [17] and explored elsewhere in single-user IEC applications [18]. In particular, in these systems images are represented by ANN-like networks called CPPNs, which are briefly reviewed in the next paragraph (a more detailed introduction is given by Stanley [18]).

Importantly for the second experiment in this domain, while each node in a traditional ANN is typically the same sigmoid activation function, each node in a CPPN has an activation function selected from a fixed *set* of such functions; the motivation for such a set is to enable more interesting visual patterns through deliberate choice of included functions. For example, sinusoidal functions may be included to enable repetitive patterns and Gaussian functions may be included to enable symmetric ones. In this domain, a CPPN is mapped to the image it represents in the following way: For each pixel in an image, the CPPN's inputs are set to its scaled Cartesian coordinates, and the output of the network is interpreted as a grayscale pixel value. In effect, the CPPN thus represents a pattern over a coordinate space, which in this case is interpreted as a picture.

The first experiment in this image evolution domain, which is described next, explores methods for seeding newly-created collaborative IEC websites with content.

Experiment 1: Bootstrapping Collaborative IEC The goal of this experiment is to show that IEC+HCM can be applied to evolve aesthetic images through selection purchased from a diffuse cloud of users. One practical application of such a technique is to *bootstrap* newly launched collaborative IEC websites with initial content. For example, for a site like Picbreeder to attract users, it helps to first have a diversity of aesthetic images that users can explore and interact with. Problematically, such initial content is difficult to produce automatically, because aesthetic evaluation generally requires human judgment. On the other hand, it is laborious and uninteresting for humans to generate because initially random images are of poor quality. Therefore, paid users are instrumental for generating such initial content with sufficient quality.

However, because IEC+HCM incurs a financial cost for each evaluation, it becomes important to leverage human input as efficiently as possible. Thus a promising approach to increase IEC+HCM's financial viability is to first generate a diversity of content *algorithmically*. Such content is more interesting than the random genomes that would otherwise seed evolution, although still in need of further human refinement.

Thus, to investigate this idea two versions of IEC+HCM are run: one method that is first seeded with pre-evolved impressive images, and another that is instead initialized with random genomes. For the unseeded runs, evolution starts from simple random CPPNs in the same way as most other CPPN-encoded image evolution applications [17, 18]. For the seeded runs, the setup of Lehman and Stanley [11] was applied to first evolve impressive artifacts, of which the most impressive from 20 separate runs were sampled to seed evolution. Figure 3 shows the representative products of both methods.

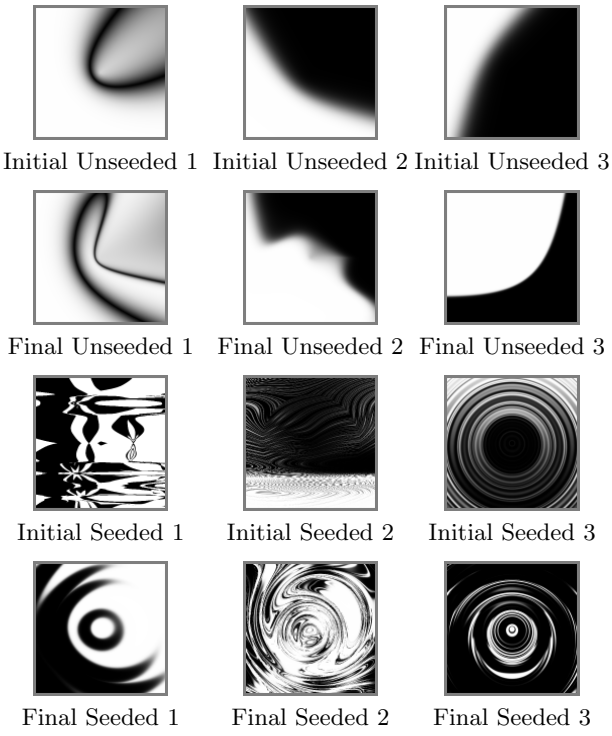

Fig. 3: **Typical Products of the Seeding Experiment.** Images from three representative runs of the seeded and unseeded methods are shown. The labels indicate whether the images are the most-preferred images from the initial or the final generation, and whether they are from the seeded or unseeded method; the number distinguishes separate runs. The main results are (1) that there is a large difference in complexity and quality between the unseeded and seeded runs, and (2) that for the seeded runs there is a noticeable divergence between the initially-preferred seed and the final most-preferred evolved image.

Because judging aesthetic appeal requires subjective human evaluation, AMT was also applied to investigate the products of the two methods. In particular,

runs of each method were first arbitrarily paired for comparison. Then, the most preferred images from the initial and final generations of both methods were placed in random order and uploaded to the same AMT evaluation task used for IEC, but with a larger number of separate user evaluations (ten instead of five).

The results of this evaluation process (seen in figure 4) show that the most preferred "impressive" seed images from the first generation of the seeded runs are rated significantly more aesthetically pleasing than are the first generation images from the unseeded runs (Mann-Whitney U-test; $p < 0.05$), supporting their motivation. Furthermore, the champion of the final generation of the seeded runs (i.e. the most-preferred product of human elaboration of the seed images) is rated significantly more pleasing than both the initial generation of the seeded runs and the final generation of the unseeded runs ($p < 0.05$). In this way, the results support the hypotheses that IEC+HCM can be leveraged to evolve increasingly aesthetically pleasing artifacts and that seeding IEC+HCM with pre-evolved artifacts can more efficiently leverage human evaluations. Thus seeded IEC+HCM may prove a viable technique for bootstrapping collaborative IEC websites.

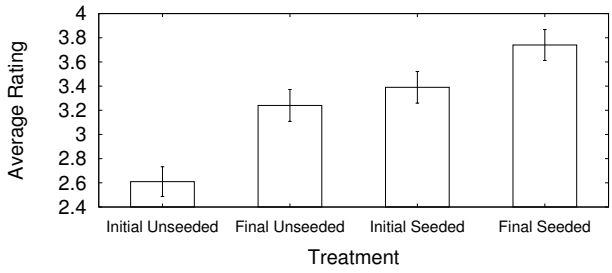

Fig. 4: **Seeding experiment evaluation.** An independent evaluation comparing the champions of the first and final generations of both the unseeded and seeded methods is shown. The main result is that the final seeded champions are on average rated significantly more aesthetically pleasing than both the final unseeded and initial seeded images.

Experiment 2: Validating Components of an IEC System The next experiment is motivated by the desire to make principled design decisions while creating a collaborative IEC website or single-user IEC application. That is, it is difficult for a system designer to decide objectively on appropriate parameter settings or what genetic encoding is best, especially when the quality of such decisions depends upon subjective factors aggregated across all targeted users (e.g. the aesthetic quality of artifacts evolved under such a decision). The problem is that the most readily available heuristic for the designer, i.e. his own aesthetic

preferences or those of his team, may not well reflect the broader preferences of the average target user.

While single-user IEC applications are easy to revise from user feedback even after they have been first released, launching an IEC website inherently involves a certain level of commitment to the domain. That is, changing the domain after the website has launched may invalidate already-evolved content, potentially alienating users whose creative products are deleted. Additionally, if previous content cannot be merged into the system after a domain or encoding changes, then its loss will make the site as a whole less engaging. Therefore it is desirable to avoid such problems and launch a better initial product.

A potential solution is to run controlled experiments with IEC+HCM to collect empirical evidence of a change's impact from a representative sample of potential users. That is, the quality of results from IEC with different parameter settings or features can be compared, by paying users through AMT to perform selection and then by paying other users to compare the final results. In this way, the particular preferences of the system's designers need not dominate the design of the system itself.

As a simple example, the second experiment evaluates the hypothesis that the additional activation functions of CPPNs improve the aesthetic quality of CPPN-evolved images beyond the use of simpler ANNs [18, 17]. A third version of the image evolution task was devised with simple ANNs (i.e. standard ANNs with a single sigmoid activation function) substituted for CPPNs (which have an extended set of activation functions). Only the activation function set is varied, all other aspects of the encoding remain unchanged. In this way, the aesthetic quality of products evolved with CPPNs could be compared to those evolved with simpler ANNs. Furthermore, taking into account the advantages of seeded IEC+HCM runs demonstrated in the previous experiment, only a seeded method with simple ANNs was considered. Note that what is impressive or rare in a particular genetic space depends upon the encoding; therefore, to accomplish seeding with these ANNs required evolving impressive artifacts in this different genetic space. Thus, ten additional IEC+HCM runs were conducted, with ANNs seeded with impressive ANNs in the same way as in the previous experiment with CPPNs.

The effect of replacing CPPNs with ANNs on the results of IEC+HCM is shown in figure 5. As expected, these images differ noticeably from the previous results with CPPNs shown in figure 3. An empirical investigation of the aesthetic difference between the seeded IEC+HCM methods with CPPN and ANNs was then conducted similarly to the previous experiment: AMT users compare the products from paired runs of this simple ANN method (shown in figure 5) and the previous IEC+HCM method with CPPNs (i.e. the final seeded images from figure 3). Figure 6 shows the results: Expanding the set of activation functions in CPPNs facilitates evolving more aesthetically pleasing images. This result demonstrates how IEC+HCM enables objective investigations of the impact of different system features.

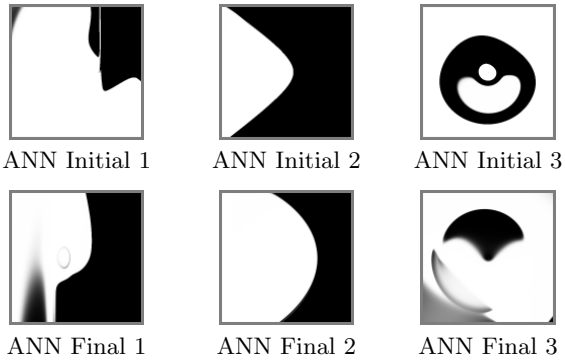

Fig. 5: **Typical Products of the Seeded ANN Runs.** Images are shown from three representative runs of seeded IEC+HCM with ANNs (instead of with CPPNs as in the previous experiment). The qualitative difference between these images and those evolved with CPPNs (figure 3) suggests that the added activation functions of CPPNs impact the kind of images likely to be evolved.

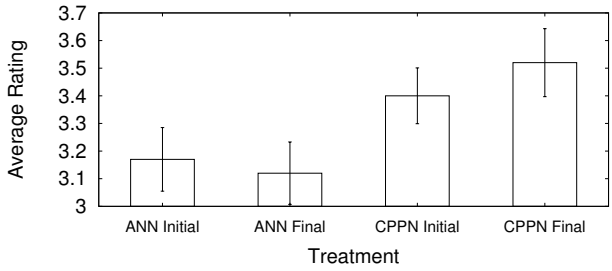

Fig. 6: **Feature Validation Experiment Evaluation.** An independent evaluation comparing the initial generation and final generation champions of seeded runs with the ANN method (ANN initial and ANN final) with those with of the CPPN method (CPPN initial and CPPN final) are shown. The main result is that the final generation CPPN images are judged significantly more aesthetically pleasing than either of the two classes of ANN images (Mann-Whitney U-test; $p < 0.05$). The conclusion is that CPPNs facilitate evolving more aesthetic images than ANNs.

4.3 Experiment 3: Evolving Image Layouts with IEC+HCM

The third experiment investigates whether IEC+HCM can expand the range of domains where large-scale IEC can be effectively applied. While large-scale IEC is currently applicable only to domains that are sufficiently enjoyable to attract volunteer users, the IEC+HCM approach can potentially be applied to any domain regardless of how engaging it is, and can be scaled to the extent that funds are available to do so.

Interestingly, the previous experiments with evolving images provide tentative evidence for this hypothesis: The IEC+HCM setup still produced aesthetic improvements in evolved images even though it perverts what is typically enjoyable about such domains, i.e. as a means for expressing personal creativity [19, 17]. That is, while in most IEC image evolution systems the user intentionally and directly drives the creative process through selection, with IEC+HCM a particular human's input is only a transient interchangeable part of a larger process, diluting any overarching influence on evolution. Supporting this idea, of the 620 unique AMT users who contributed to the image evolution IEC+HCM experiments, 419 users completed only *one* evaluation task (i.e. they interacted with the system only once and could not have seen any effect of their influence), and only 65 users contributed more than four evaluations. Thus it is unlikely that any particular user will receive the satisfaction of seeing their aesthetic influence realized.

However, image evolution still offers the potential of novelty between evaluations, which may be interesting for a user even if IEC+HCM does not allow for directly expressing creativity as in other IEC systems. Thus to more directly test the hypothesis that IEC+HCM can extend the reach of large-scale IEC to domains not inherently enjoyable, the third experiment explores an intuitively less enjoyable task, that of evolving the layout of an image composition. In particular, the task is to evolve the relative positions of a fixed set of images (seen in figure 7) to maximize the aesthetic appeal of the composition. Unlike the image evolution domain, the potential for novelty is limited because the components of the image are always the same and uninteresting.

The domain and genetic encoding are illustrated by figure 7. Note that the same IEC+HCM setup as previously described was adapted for this third experiment but with only a single unseeded method. While seeding with impressive pre-evolved layouts might accelerate progress in this domain, such seeding is not necessary to verify the hypothesis.

The products of this experiment are shown in figure 8. The results were validated similarly to the previous experiments, by presenting them to be rated by a larger set of AMT users. However, instead of comparing between methods, evolved artifacts are compared over generations. The idea is to demonstrate that progress in aesthetic evolution is occurring. The aggregated ratings from the larger validation evaluation are shown in figure 9. As expected, the most-preferred layouts from the final generation are rated significantly more pleasing in appearance than those from the first generation, thus supporting the conclusion that evolutionary progress was facilitated by IEC+HCM in this domain.

Fig. 7: **The Image Layout Domain.** The image layout experiment evolves a composition of the four shown images through IEC+HCM. The encoding is a simple list of Cartesian coordinates that specify the offset of each of each image. Initial genomes are generated such that they cluster the four images near the upper left corner, providing a predictably poor starting arrangement for human-guided evolution to improve upon. Mutation perturbs the coordinates of one image out of the four, adding to the x and y coordinate a separately number chosen uniformly between -50 and 50.

Note that while the domain itself is somewhat trivial, the results provide an existence proof that IEC+HCM can extend large-scale IEC to domains that are not inherently enjoyable.

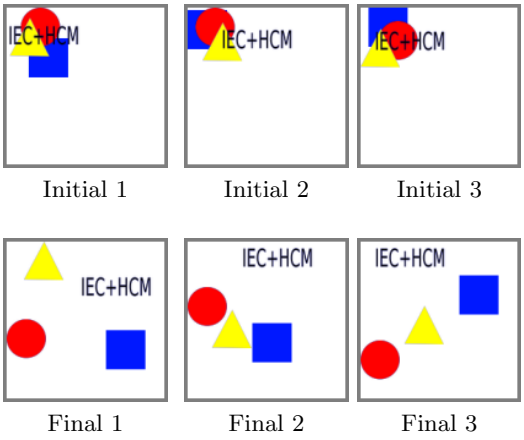

Fig. 8: **Typical Products of the Layout Evolution Experiments.** Images are shown from three representative runs of IEC+HCM in the layout evolution domain. In particular, the most-preferred image from the initial and final generation of the runs are shown. Over evolution, the images composing the layouts expand to better fill the space. The conclusion is that IEC+HCM can be successfully applied even in domains that are not inherently enjoyable.

5 Discussion and Future Work

This paper investigated leveraging markets for human computation to support large-scale IEC in three ways. Exploratory experiments in this paper showcase

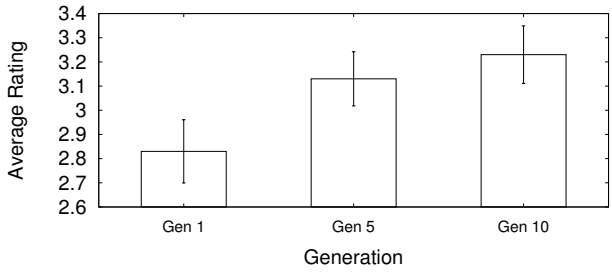

Fig. 9: **Image Layout Experiment Evaluation.** An independent evaluation comparing the champions of the first, fifth, and final (tenth) generations from the ten independent IEC+HCM runs in the image layout domain is shown. The main result is that the image layout of final champions is judged significantly more aesthetic than that of the first generation (Mann-Whitney U-test; $p < 0.05$).

how the ability to pay for human computation potentially can bootstrap IEC websites, inform the design of such websites or single-user IEC systems, and act as a viable alternative to such websites when the domain is not enjoyable or engaging.

In this way, an interesting advantage of the IEC+HCM approach is that it bypasses the significant problem of user fatigue in IEC [19] through paying users directly, without constraining the domain. Of course, the trade-off is that pairing IEC with human computation incurs an explicit financial cost per evaluation. While such financial costs have always been implicit in strictly computational EC (e.g. costs to maintain a cluster of computers necessary for large-scale experimentation) and collaborative IEC websites (e.g. server costs), the price of human computation does not similarly benefit from Moore's law. Thus large-scale IEC+HCM may be most applicable for unengaging domains limited by difficulty in applying appropriate selection pressure, and also possibly for commercial applications where the cost of IEC+HCM is less than the value of the evolved artifact.

So while the IEC+HCM mechanism can be leveraged to improve the design and engagement of single-user IEC systems and collaborative IEC websites, its most interesting implication may be that exploiting it on a large scale may potentially lead to results exceeding current approaches in evolutionary robotics or artificial life. That is, to the extent that current approaches are limited by lack of appropriate selection pressure [25, 14, 10], and to the extent that human judgment can remedy such limitations [5, 22], human computation may be a technique that can be exploited to further the state of the art in EC. For example, large-scale IEC+HCM with a significant budget applied to evolving virtual creatures might produce creatures with complexity and functionality beyond the reach of current methods. In this way, an interesting direction for further experimentation is to apply IEC+HCM to evolve controllers for evolutionary robotics or artificial life experiments.

6 Conclusion

This paper explored combining interactive evolution with human computation markets to purchase a powerful form of selection pressure. The promise of the approach was shown in preliminary experiments evolving aesthetic images and the layout of image compositions. Applying the same techniques in other domains limited by lack of appropriate selection pressure may enable evolution of more complex artifacts or behaviors than previously possible. The conclusion is that human computation markets may be an important tool for supporting collaborative IEC websites as well as for extending the reach of large-scale IEC beyond only task domains that are enjoyable.

References

1. Chou, C., Kimbrough, S., Sullivan-Fedock, J., Woodard, C., Murphy, F.: Using interactive evolutionary computation (iec) with validated surrogate fitness functions for redistricting. In: Proceedings of the fourteenth international conference on Genetic and evolutionary computation conference. pp. 1071–1078. ACM (2012)
2. Clune, J., Lipson, H.: Evolving three-dimensional objects with a generative encoding inspired by developmental biology. In: Proceedings of the European Conference on Artificial Life (2011)
3. Den Heijer, E., Eiben, A.: Comparing aesthetic measures for evolutionary art. Applications of Evolutionary Computation pp. 311–320 (2010)
4. Goldberg, D.E.: Simple genetic algorithms and the minimal deceptive problem. In: Davis, L.D. (ed.) Genetic Algorithms and Simulated Annealing, Re- search Notes in Artificial Intelligence. Morgan Kaufmann (1987)
5. Gruau, F., Quatramaran, K.: Cellular encoding for interactive evolutionary robotics. In: Fourth European Conference on Artificial Life. pp. 368–377. MIT Press (1997)
6. Hastings, E., Stanley, K.: Interactive genetic engineering of evolved video game content. In: Proceedings of the 2010 Workshop on Procedural Content Generation in Games. p. 8. ACM (2010)
7. Ipeirotis, P.: Analyzing the amazon mechanical turk marketplace. XRDS: Crossroads, The ACM Magazine for Students 17(2), 16–21 (2010)
8. Kittur, A., Chi, E.H., Suh, B.: Crowdsourcing user studies with mechanical turk. In: Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems. pp. 453–456. ACM (2008)
9. Kosorukoff, A.: Human based genetic algorithm. In: Systems, Man, and Cybernetics, 2001 IEEE International Conference on. vol. 5, pp. 3464–3469. IEEE (2001)
10. Lehman, J., Stanley, K.O.: Abandoning objectives: Evolution through the search for novelty alone. Evolutionary Computation 19(2), 189–223 (2011)
11. Lehman, J., Stanley, K.O.: Beyond open-endedness: Quantifying impressiveness. In: Proceedings of Artificial Life Thirteen (ALIFE XIII) (2012)
12. MacCallum, R., Mauch, M., Burt, A., Leroi, A.: Evolution of music by public choice. Proceedings of the National Academy of Sciences 109(30), 12081–12086 (2012)
13. McCormack, J.: Open problems in evolutionary music and art. Applications of Evolutionary Computing pp. 428–436 (2005)

14. Miconi, T., Channon, A.: An improved system for artificial creatures evolution. In: Proceedings of the Tenth International Conference on Artificial Life (ALIFE X). pp. 255–261. MIT Press (2006)
15. Orkin, J., Roy, D.: The restaurant game: Learning social behavior and language from thousands of players online. Journal of Game Development 3(1), 39–60 (2007)
16. Risi, S., Lehman, J., D'Ambrosio, D.B., Hall, R., Stanley, K.O.: Combining search-based procedural content generation and social gaming in the petalz video game. In: Proceedings of the Artificial Intelligence and Interactive Digital Entertainment Conference (AIIDE 2012) (2012)
17. Secretan, J., Beato, N., D'Ambrosio, D., Rodriguez, A., Campbell, A., Folsom-Kovarik, J., Stanley, K.: Picbreeder: A case study in collaborative evolutionary exploration of design space. Evol. Comp. (2011), to appear.
18. Stanley, K.: Compositional pattern producing networks: A novel abstraction of development. Genetic Programming and Evolvable Machines 8(2), 131–162 (2007)
19. Takagi, H.: Interactive evolutionary computation: Fusion of the capacities of EC optimization and human evaluation. Proceedings of the IEEE 89(9), 1275–1296 (2001)
20. Von Ahn, L.: Games with a purpose. Computer 39(6), 92–94 (2006)
21. Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., Blum, M.: recaptcha: Human-based character recognition via web security measures. Science 321(5895), 1465–1468 (2008)
22. Woolley, B.G., Stanley, K.O.: Exploring promising stepping stones by combining novelty search with interactive evolution. CoRR abs/1207.6682 (2012)
23. Yu, L., Nickerson, J.: An internet-scale idea generation system. ACM Transactions on interactive Intelligent Systems 3(1), 2013 (2011)
24. Yu, L., Nickerson, J.V.: Cooks or cobblers?: crowd creativity through combination. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1393–1402. ACM (2011)
25. Zaera, N., Cliff, D., Bruten, J.: (Not) evolving collective behaviours in synthetic fish. In: From Animals to Animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior. MIT Press Bradford Books. (1996)