Searching for Good and Diverse Game Levels

Mike Preuss
European Research Center
for Information Systems
WWU Münster, Germany
mike.preuss@uni-muenster.de

Antonios Liapis
Center for Computer Games Research
IT University of Copenhagen
Copenhagen, Denmark
anli@itu.dk

Julian Togelius
Center for Computer Games Research
IT University of Copenhagen
Copenhagen, Denmark
julian@togelius.com

***Abstract*—In procedural content generation, one is often interested in generating a large number of artifacts that are not only of high quality but also diverse, in terms of gameplay, visual impression or some other criterion. We investigate several search-based approaches to creating good and diverse game content, in particular approaches based on evolution strategies with or without diversity preservation mechanisms, novelty search and random search. The content domain is game levels, more precisely map sketches for strategy games, which are meant to be used as suggestions in the Sentient Sketchbook design tool. Several diversity metrics are possible for this type of content: we investigate tile-based, objective-based and visual impression distance. We find that evolution with diversity preservation mechanisms can produce both good and diverse content, but only when using appropriate distance measures. Reversely, we can draw conclusions about the suitability of these distance measures for the domain from the comparison of diversity preserving versus blind restart evolutionary algorithms.**

I. Introduction

In procedural content generation (PCG), algorithms are used to create game content such as levels, rules, items or characters. There are several good reasons for developing and using PCG solutions, such as reducing the cost of game development, adapting games to individual players, enabling infinite games and computationally studying game design. There are also several desirable properties of a PCG algorithm: that it produces high-quality content (on average or in the best case), that it is reliable (never produces bad content), that it is fast, that the generated content is believable and that it is diverse. In most cases, a single PCG algorithm cannot exhibit all of these desirable properties, and therefore some sort of tradeoff needs to be made [1].

In this paper, we will focus on two of those properties: (best case) quality and diversity. There are many PCG problems where these are the most important requirements of the algorithm. Consider the case when the algorithm delivers solutions for a human to choose from, such as when content generation interacts with a human designer in a mixed-initiative process. Here, it is important to deliver candidate content artifacts that the human will consider worth examining and building on (best case quality), and that are sufficiently different from each other and from what the human is producing him- or herself (diversity). This can be contrasted with many runtime content generation tasks, where speed and average or worst-case quality are the most important requirements. The current paper considers the concrete case of mixed-initiative map sketch creation for strategy games, and investigates search-based methods for generating good and diverse map sketches.

In order to do this, we need working definitions of "good" and "diverse". We will take for granted that there are several different aspects of game content quality, which may be partially conflicting. In previous work we have investigated the various qualities of maps that can be measured specifically for strategy games, and explored the use of multiobjective evolution to balance the conflicts between these metrics [2]. We have also defined a set of quality metrics for map sketches that can be applied across a number of related types of game content, including strategy game maps, first-person shooter levels, roguelike dungeons and platform game levels [3]. These quality metrics are based on concepts of area control, exploration and balance, and will be used as objectives in the experiments in this paper. We will assume that a user cares about all or a subset of these metrics, and will be searching for maps that maximize the sum of the chosen metrics.

The definition of "diverse" is no less complicated than the definition of "good". Essentially, the question is in what way the game content artifacts are different from each other: should they look different, play differently, be structurally different or differ in some other way? In this paper, we will consider three different measures of diversity for map sketches. While the details will be discussed later, we will base our diversity measures on the following concepts:

- *Tile-based* diversity. This can also be called microscopic or genotypic diversity. Here, difference is measured in the most low-level way, as the number of map tiles that differ from one map sketch to another.
- *Objective-based* or quality-based diversity. Here, we measure diversity in terms of the same quality metrics we use for searching for good maps.
- *Visual impression* diversity. Here, we extract visual features of the maps through metrics related to balance or grouping, and measure diversity along those metrics.

Certain diversity measures among those proposed in this paper are likely to appeal more to specific human users, although describing diversity can also be highly subjective and depend on the task at hand (e.g. playing a game level versus designing one). This paper, however, will also investigate how the different diversity measures can affect search algorithms with diversity preservation. The three types of search/optimization algorithms we will use are:

- Traditional objective-driven search, as performed by most evolutionary algorithms (with restarts, this can also be used to generate a diverse solution set).

- Objective-driven search with diversity preservation measures, as in niching evolutionary algorithms.
- Diversity-only search, such as novelty search.

Note that we may count all of these methods as *multimodal optimization* methods (at least for constrained search spaces, this is also the case for novelty search), but only the second method type actually contains niching algorithms. However, these terms are more often used for problems with numerical representations.

A. Innovations of this paper

This paper aims to discover effective and efficient ways of generating game levels that are both good and diverse. The main concrete outcome of the paper is an effective and efficient method for generating map sketches as suggestions of the Sentient Sketchbook mixed-initiative game level design tool. In doing so, the impact of a distance heuristic is tested on different evolutionary algorithms which take into account the divergence of a population, namely niching methods and novelty search methods. The notion of "diversity" in game content is explored via alternatives to the previously explored tile-based diversity metric [4], [5], with several important findings on how the distance metric affects both the appearance and functionality of generated content as well as the performance of the evolutionary methods. Moreover, this is the first elaborate application of niching methods in non real-valued search spaces and several insights on the importance of distance characterizations on their performance is gleaned, along with the impact of constrained search spaces.

B. Structure of this paper

Section II describes Sentient Sketchbook, a mixed-initiative game level design tool which generates map sketches as suggestions to the human user. The appearance and functionality of these suggestions is, in essence, the direct motivation for this paper. Section II also discusses the six quality metrics for game levels developed as part of the Sentient Sketchbook project. Section III describes the three different distance measures which were developed for map sketches, two of which are new for this paper. Section IV presents the different algorithms we will use for searching for good and diverse map sketches. Section V reports on systematic experiments with all three algorithms and diversity measures. In the discussion of Section VI we return to the question of defining — and searching for — good and diverse game content, informed by the results of the experiments, and provide directions for future work. Conclusions are drawn in Section VII.

II. Sentient Sketchbook

Sentient Sketchbook is a computer-aided design tool which allows a human designer to create game levels while a computational designer suggests alternatives to the user's creations [4]. By focusing designer effort on the rough outline (sketch) of a game level, Sentient Sketchbook allows for rapid prototyping and concept development. Several automated features, such as the evaluation of game level quality and the generation of alternative designs in real-time, allow for a dialogue between the human user and the machine intelligence

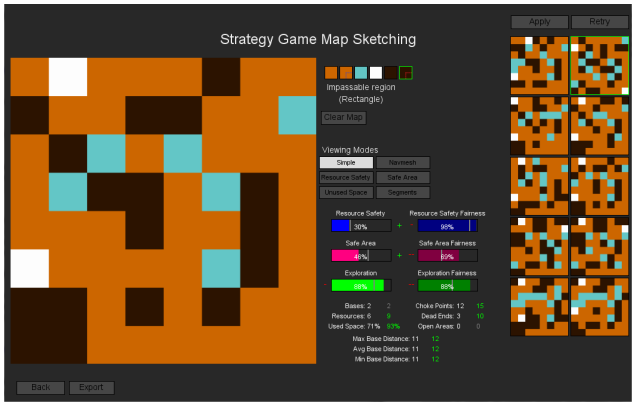
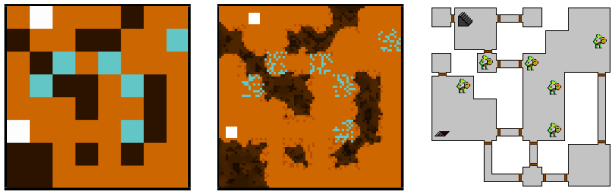


Fig. 1: The user interface of Sentient Sketchbook, as the human designer edits their map sketch on the canvas to the left while the computational designer suggests alternatives in the far right edge of the editing window.



(a) Map Sketch (b) Final Strategy Game Level (c) Final Dungeon Level

Fig. 2: A map sketch (Fig. 2a) and the strategy game level it creates (Fig. 2b): white tiles are bases, cyan tiles are resources and dark tiles are impassable. The map sketch can also be envisioned as a dungeon (Fig. 2c), where white tiles are entrances or exits and cyan tiles are monsters.

which is expected to promote mixed-initiative co-creativity [6]. While several types of levels can be created via Sentient Sketchbook [3], this paper will focus on strategy game levels.

A. Representation

Sentient Sketchbook operates on abstractions of game levels represented as *map sketches*. These map sketches consist of low-resolution versions of a complete strategy game level set on a small grid of tiles. Tiles of the map sketch can be impassable (blocking unit movement), resources, bases and empty; the latter three tile types are passable, i.e. allow unit movement. Players are assumed to start from a base tile and collect resources in order to build units which travel along passable paths to attack enemy bases. Due to the small size of the map sketches and the few tile types, the map sketch is stored directly in the genotype. The genotype consists of an array of integers, each integer corresponding to the type of a single tile on the map sketch.

B. Suggestions

A significant innovation of Sentient Sketchbook is the presentation to the user of computer-generated alternatives of their current sketch. These computational *suggestions* are generated via genetic algorithms running on short evolutionary

sprints and shown in real-time as the users edit their own map sketch. Evolution begins from an initial population consisting of mutations of the user's current sketch; this results in some structural similarity between computational suggestions and human creations, so that the suggestions are clearly connected to the sketch that is being worked on. In order to ensure that the shown suggestions represent playable game levels, a constraint that all bases and resources must be connected via passable paths is enforced.

Sentient Sketchbook can present up to 12 suggestions to the user at any time. In the previous version of Sentient Sketchbook described in [4], six suggestions target strategic quality and six suggestions target visual diversity. The former six suggestions are generated via six separate threads of constrained evolution via the FI-2pop genetic algorithm [7]; each thread targets one of the strategic properties in Section II-C. The latter six suggestions are generated via constrained novelty search [5] which targets tile-based diversity; a description of novelty search and tile-based diversity will be provided in Sections IV-A and III-A respectively. The previous version of Sentient Sketchbook makes a distinction between "good" and "diverse" suggestions by evolving them via different methods; no assumptions are made that the diverse suggestions are good and vice versa. Additionally, suggestions generated via objective-driven search explicitly target a single objective rather than all of them. In contrast, this paper explores methods for generating good *and* diverse levels, with "good" evaluated on a combination of all strategic qualities and "diverse" evaluated on visual or functional differences.

C. Strategic Qualities

Map sketches are evaluated on six strategic qualities, inspired by game design patterns of area control, exploration and balance [8] and verified by game developers. The six fitness dimensions are: f_{res}, which evaluates how safe (i.e. nearby) resources are to any base and b_{res} how balanced the distribution of safe resources is among bases; f_{saf}, which evaluates how many safe passable tiles are near any base and b_{saf} how balanced this distribution of safe passable tiles is among bases; f_{exp}, which evaluates how much exploration is needed to find all bases starting from all other bases and b_{exp} how balanced such exploration is when starting from different bases. Arguments for and mathematical definitions of these fitness dimensions are included in [3].

Note that only *feasible* maps can be evaluated for their strategic qualities at all. Maps which lack possible paths between some bases are *infeasible* as they cannot be played on, and therefore have no computable qualities. The larger the map dimensions, the more of the map space is populated by infeasible maps.

III. Distance measures

It is not straightforward to describe how different two map sketches are, and thus assessing whether generated (or hand-crafted) levels are "diverse" will be affected by this distance description. Moreover, a quantifiable measure of distance can greatly affect both novelty search and niching for promising solutions. This paper explores different measures of diversity based on direct image comparisons, based on informed measures of strategic quality or based on measures of visual impression inspired by studies on human cognition.

A. Tile-based distance

The most straightforward distance between two map sketches is based on their representation as 2D images, and can be derived by comparing them on a per-pixel basis. Tile-based distance enumerates the number of tiles (pixels) which have a different tile type (color) from one sketch to the other. Eq. (1) describes this simple measure of diversity.

$$d_{tile}(i,j) = \frac{1}{wh}\sum_{x=1}^{w}\sum_{y=1}^{h} P_{x,y}(i,j) \quad (1)$$

where w and h the width and height of the sketch respectively; $P_{x,y}(i,j)$ is 0 if the tile of map i at position (x,y) is of the same type as that of map j (at the same position), and 1 otherwise.

B. Objective-based Distance

Since map sketches are not mere images but represent a strategy game level with specific affordances, mechanics and game properties, another description of distance between two map sketches can be based on the strategic qualities which have been defined in Section II-C. Treating the six objectives of Section II-C as a vector, the objective-based distance calculates the Euclidean distance between the two maps' objective scores, represented as a vector $\vec{O} = \langle f_{res}, f_{saf}, f_{exp}, b_{res}, b_{saf}, b_{exp}\rangle$.

C. Visual Impression Distance

The tile-based distance of eq. (1) is a straightforward, context-independent measure of visual diversity which has certain desirable qualities, such as being orthogonal to the objectives being optimized via niching. However, tile-based distance does not account for certain fundamental properties of human perception which bias the way humans appreciate 2D artifacts. Inspired by cognitive psychology and the notion of perceptual forces [9], we define visual-based objectives which correspond to a map's balance (or symmetry), grouping and concentration, i.e. how balanced two halves of the map sketch are, how grouped together elements such as impassable tiles are and what proportion of tiles of one type exist in each half of the map sketch. Similar to objective-based distance, these dimensions of visual impression are treated as a vector $\vec{V}$ and visual impression distance amounts to a Euclidean distance between the two maps' scores in these visual impression dimensions. The 20 separate dimensions of visual impression included in $\vec{V}$ (preceded by their index ranges) are:

- 1-2 Vertical Symmetry: evaluates whether the left half is an exact reflection of the right half. Similarly for Horizontal Symmetry (bottom half and top half).
- 3-8 Vertical Impassable (Resource, Base) Balance: evaluates how close the number of impassable (resource, base) tiles in the left half is to their number of the right half. Similarly for Horizontal Balance (bottom half and top half).
- 9-14 Left Half Concentration of Impassable (Resource, Base) Tiles: evaluates which portion of the total number of impassable (resource, base) tiles is in the left half. Similarly for Top Half Concentration.

15-17 Diagonal Concentration of Impassable (Resource, Base) Tiles: evaluates which portion of the total number of impassable (resource, base) tiles is in the top left and bottom right quadrants.

18 Impassable Ratio: evaluates which portion of the map's tiles are impassable.

19 Impassable Segments: evaluates how many distinct segments (via 4-directional flood fill) are formed by the impassable tiles, compared to all impassable tiles.

20 Largest Segment: evaluates which portion of the map's impassable tiles are connected in the single largest impassable segment.

IV. Optimization Algorithms

Searching for several good *and* diverse solutions at the same time is the main goal of the emerging area of *multimodal optimization* (MMO), of which [10] provides a recent survey. Niching is one of the standard techniques and its origins go back to the 1970s. However, it is not the only possibility. One may also disregard objectives in the first place with novelty search and try to detect good solutions among the returned set. Additionally, the CEC 2013 niching competition [11] has shown that simply applying a restart EA and collecting the detected local optima is a good default (non-niching) technique. Recent research on quality indicators for MMO [12] reveals that a) there are many similarities between multimodal and multi-objective optimization, and b) that measuring the quality of MMO algorithms is not at all trivial. However, most of the MMO research has concentrated on real-valued search spaces, and for "non-standard" representations, many aspects of MMO remain unclear.

Due to the low ratio of feasible solutions among randomly generated ones of the tackled optimization problem, all of the following algorithms employ the "feasible first" principle, i.e. feasible solutions always get precedence above infeasible ones whenever there is a choice. For novelty search, this preference of feasible solutions could be likened with Minimal Criteria Novelty Search [13]. Also note that we regard mutation as a black box component provided by the problem code. It changes the map, but we cannot control how large the change is.

A. Novelty search

Novelty search has been proposed as an alternative to objective-driven search for cases where the objective function is ill-defined, deceptive, subjective or unknown [14]. The selection process of novelty search favors those individuals which are different from the remaining population as well as from past solutions, which are stored in a novel archive. Novelty search ranks individuals according to their novelty score ρ of eq. (2), which corresponds to the average distance of an individual i with its closest neighbors. In this instance, the 5 most novel individuals of a generation are inserted in the novel archive, while the number of closest neighbors (k) when calculating ρ is 20.

$$\rho(i) = \frac{1}{k} \sum_{j=1}^{k} d(i, \mu_j) \tag{2}$$

where μ_j is the j-th-nearest neighbor of i (within the population and in the archive of novel individuals); distance $d(i, j)$ is a domain-dependent metric which evaluates the "difference" between individuals i and j.

Due to the constrained space, Minimal Criteria Novelty Search [13] is implemented in the sense that infeasible offspring are given the lowest preference. Unlike other evolutionary strategies in this paper, novelty search uses (X+X) selection scheme for a population size of X, i.e. each parent in the population produces an offspring and the novelty score of each individual in the population (2X individuals) are evaluated according to eq. (2); the X most novel (also feasible) individuals are chosen to replace the current generation, and the 5 most novel among those are inserted into the novel archive.

B. Restart evolution strategies

Our search space consists of a matrix of nominal scaled variables, so that highly specialized algorithms such as the *covariance matrix adaptation evolution strategy* (CMA-ES) [15] are not applicable. However, we can still borrow some of its sophisticated restart criteria and use them within a standard ES [16], run this up to a predefined budget and collect its single solutions. Note that the main difference between this approach and the niching method described in the following Section lies in the missing restart location determination: we just restart at a random location. Comparing this approach to a niching algorithm may be seen as the *experimentum crucis* for the latter: if they perform similarly, the restart "organization" (and by that basically the niching component) does not work. In our context, it is highly likely to produce an infeasible solution by mutating an existing one. We therefore employ a plus-selection, meaning that older individuals survive indefinitely if they are better than their offspring.

C. NEA2

The *niching evolutionary algorithm 2* (NEA2) won the CEC 2013 niching competition [11] and may also be applied in non real-valued search spaces as it solely relies on a suitable distance definition between single solutions and a ranking of their qualities. It has been introduced in [17] and consists of two phases that are repeated until the algorithm is stopped: in the initial phase, a relatively large population (here: 100 random samples) is employed for determining basins of attraction by means of the *nearest-better clustering* algorithm (NBC, [18]). Based on a distance matrix, we connect every solution to the nearest one within the sample that is better. Then, we cut the longest connections according to a heuristic rule and thereby obtain clusters. The default rule cuts all connections that are larger than 2 times the average connection length; it is also applied here. In the second phase, a suitable local optimizer (here: the EA described in sec. IV-B) is run in every identified basin until it stagnates, and the approximated local optima are put into an archive. If the budget is not yet used up, we start again with phase 1, leaving the archive intact. Finally, the archived solutions are the result set.

V. Experimental Analysis

We know that in real-valued search spaces, it is possible to provide a set of solutions that is at the same time good and

TABLE I: Small maps: minimal neighbor distances between the chosen 6 solutions, averaged across 21 runs. Rows stand for the measured values in each diversity metric (averaged from 21 repeats), columns give the diversity employed for the optimization. VI stands for visual impression distance. The 4 algorithms are NEA2 (with 6 distance criteria), rES is the restart evolution strategy, novelty search (with the same 6 distance criteria) and random search.

run	NEA 2						rES	Novelty search						random
diversity	tile	object.	VI-all	VI-bal	VI-grup	VI-con		tile	object.	VI-all	VI-bal	VI-grup	VI-con	
tile	0.555	0.565	0.555	0.558	0.562	0.555	0.557	0.563	0.544	0.549	0.545	0.559	0.567	0.550
objective	0.084	0.081	0.099	0.093	0.084	0.096	0.086	0.151	0.146	0.155	0.158	0.148	0.149	0.156
VI-all	0.239	0.239	0.265	0.273	0.253	0.259	0.257	0.212	0.210	0.208	0.204	0.223	0.220	0.257
VI-bal	0.238	0.245	0.288	0.303	0.261	0.266	0.250	0.226	0.226	0.216	0.229	0.219	0.226	0.245
VI-grup	0.094	0.094	0.095	0.103	0.096	0.108	0.094	0.085	0.081	0.083	0.077	0.076	0.077	0.077
VI-con	0.216	0.229	0.233	0.246	0.233	0.234	0.231	0.185	0.196	0.194	0.199	0.202	0.206	0.227

TABLE II: Large maps: minimal neighbor distances between the chosen 6 solutions averaged across 21 runs; the same format as Table I is used. Random search does not reliably produce feasible solutions on large maps.

run	NEA 2						rES	Novelty search						random
diversity	tile	object.	VI-all	VI-bal	VI-grup	VI-con		tile	object.	VI-all	VI-bal	VI-grup	VI-con	
tile	0.595	0.597	0.594	0.597	0.593	0.592	0.590	0.563	0.425	0.383	0.475	0.572	0.538	–
objective	0.073	0.074	0.074	0.071	0.075	0.073	0.069	0.047	0.047	0.042	0.044	0.044	0.047	–
VI-all	0.120	0.113	0.138	0.119	0.122	0.134	0.111	0.061	0.047	0.047	0.058	0.0578	0.046	–
VI-bal	0.117	0.109	0.130	0.112	0.116	0.132	0.109	0.076	0.056	0.053	0.058	0.057	0.060	–
VI-grup	0.040	0.040	0.044	0.038	0.043	0.040	0.040	0.036	0.032	0.038	0.040	0.039	0.036	–
VI-con	0.109	0.108	0.124	0.111	0.107	0.127	0.106	0.053	0.034	0.036	0.047	0.045	0.038	–

diverse. Depending on different factors as dimensionality and ruggedness, one would usually apply either a restart EA that is rather greedy as the CMA-ES, or a multimodal (i.e. niching) EA [17], [19]. Niching methods will not work very well for very highly multimodal and/or dimensional problems; restart local search algorithms are usually preferably in such cases.

For non-numeric search spaces, one may still apply niching concepts, as these basically rely on a "suitable" distance measure, but it is currently not clear a) under which conditions niching algorithms provide an advantage over restart local search algorithms, and b) what properties a suitable distance measure must have. Of course, one may also try a restart EA that matches the given representation. We now make the following assumption: if we compare these two algorithm types on different distance measures, then the one that leads to the largest advantage in terms of niching to restart EA algorithm performance is the most suitable distance measure for the problem. The reasoning behind this assumption is that when the niching method is actually able to detect several basins of attraction by means of the distance matrix, and this improves algorithm performance, then the distance measure itself must be suitable, in that it maps the effort to get from one to another solution to a scalar value.

Research Question(s): Can we reliably provide a set of diverse *and* very good solutions? Which distance measures are most useful for doing so?

These two questions are obviously related, we will therefore structure our experimental description into two phases: phase 1 deals with distance measure selection, in phase 2 we actually compare the algorithms and obtained maps on base of this distance measure.

Pre-experimental Planning: For the restart ES and the NEA2 niching algorithm, the most important parameters are the (local search oriented) population sizes for the single runs. After some test runs, these have been fixed to a (1+20) scheme (one parent, 20 offspring, elitist selection). Accordingly, the run lengths are set to 20000, because it very rarely happens that one sees a quality improvement after this time. The (global) niching population size for NEA2 is set to 100; different sizes seemingly have very little effect. For novelty search the population is also set to 100. In order to provide a "gradient" to feasible solutions, infeasible ones are penalized per objective (out of 6 objectives) by the approximate distance to feasibility. This is encoded in a way that ensures that feasible solutions are always better than infeasible ones. In addition to the evolutionary algorithms, random search is used to generate 20000 individuals on every run; although these individuals are far more than the final 100 individuals of novelty search or the likely fewer final individuals of restart ES and NEA 2, most random individuals are infeasible and thus ignored.

Task: To obtain a proper distance measure (phase 1) for the two distance-based algorithms (Novelty search, NEA2), we select the distance criterion that enables the best distance performance (see setup) in comparison to the restart ES results when they are actually applied (runs with different distance criteria do not count). We demand that the advantage is significant for a Wilcoxon rank-sum test (α-level 0.05).

Setup: All four algorithms (random search, restart ES, NEA2, novelty search) are run with 21 repeats on small and large maps. Small maps consist of 64 tiles and must contain 2 bases and 4-10 resources; due to the small size and small number of bases, small maps are likely to be feasible ($3{\cdot}10^4$ feasible in 10^6 randomly initialized small maps). Large maps consist of 256 tiles and must contain 2-10 bases and 4-30 resources; since it is very likely that large maps contain many bases and resources, they are very unlikely to be feasible (13 feasible in 10^6 randomly initialized large maps). For the two distance-oriented algorithms (NEA2, Novelty search), 6 different distance measures are tested: tile-based distance,

objective-based distance, visual impression (VI) distance (including all features), and visual impression distance with only balance/symmetry (indices 1-8), concentration (indices 9-17), and grouping dimensions (indices 18-20), respectively. From the returned set of best solutions, we select only the valid ones and determine 6 representatives by k-medoids ($k = 6$) clustering. As a measure of diversity we employ the average nearest neighbor distance between the 6 finally selected maps; the distance is normalized to the number of dimensions in the vector, providing values in the $[0, 1]$ range. The optimization objective is the average of the six strategic qualities discussed in Section II-C.

Phase 1 Results: Tables I and II show average next neighbor distances for small and large maps on all employed distance measures. Objective values are only considered in phase 2.

Phase 1 Observations: For the small maps, the VI-all, VI-balance, and VI-concentration distance measures work well for NEA2, while no distance is especially good for novelty search. On the large maps, NEA2 profits most from employing the VI-all distance, whereas tile-based distances are suitable for novelty search. In all scenarios, we find a high correlation between the effect of the VI-all, VI-balance and VI-concentration distances and the results of NEA2. However, novelty search usually performs worse than NEA2 or rES except for the objective-based distances, which results in a wider spread in terms of objective values (including several bad maps). As expected, random search usually produces very high diversity values regardless of the distance measure. The standard deviations for most diversity values range from 5% to 10% of the measured values.

Phase 1 Discussion: To our surprise, NEA2 as well as rES usually outperform novelty search in terms of diversity. Random search, however, stably reaches good diversity with very limited quality. With the right distance measure, NEA2 is usually even better than random search (in terms of diversity). While VI-all is always a suitable choice, the 3 VI groups alone sometimes work, sometimes not, with VI-balance and VI-concentration appearing as more reliable. We therefore conclude that the VI-all measure is in general most meaningful for detecting differences in maps and use it as the method of choice for NEA2 and Novelty search in the following phase. Wilcoxon rank sum tests between the average next neighbor distances of the NEA2 (employing VI-all distance) and rES runs ($p = 0.549$ for small maps, $p < 10^{-5}$ for large maps) indicate that the difference is significant only for the latter. We conclude that the VI-all distance works but the advantage obtained from using it probably depends on a large enough search space. For phase 2, only NEA 2 and novelty search which diversity VI-all will be considered.

Phase 2 Results: Figure 3 shows the 6 selected maps of the median run (in terms of distance) per method. Sketches of the large maps reveal similar characteristics (albeit being much more detailed) and are thus not displayed. Figure 4 compares the performances in distance and objective criteria (maximization, 1.0 is the optimal value for the latter). Table III displays average quality scores of the most diverse maps.

Phase 2 Observations: Concerning the map sketches, it is rather difficult to see the difference in objective values (the top 2 rows are clearly better on average), but NEA2 and rES maps look structurally more different, including some rather weird

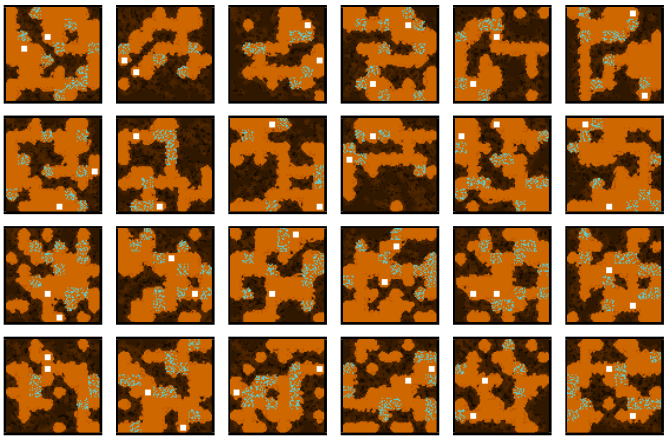


Fig. 3: Median run small maps computed by (from top to bottom) NEA2, rES, Novelty search, and random search. Mean distance values among the 6 selected maps are 0.343, 0.337, 0.300, and 0.333 respectively, per method. Mean objective values are 0.855, 0.847, 0.543, and 0.564. Generated sketches are rendered as full strategy maps for easier visual recognition.

TABLE III: Average quality scores of the 6 most diverse maps according to VI-all (maximization, 1.0=optimal).

	Small maps				Large maps			
	NEA 2	rES	nov.	rand.	NEA 2	rES	nov.	rand.
fitness	0.847	0.854	0.558	0.558	0.622	0.642	0.559	–
f_{res}	0.562	0.570	0.315	0.320	0.348	0.374	0.308	–
f_{saf}	0.782	0.800	0.373	0.380	0.469	0.512	0.376	–
f_{exp}	0.824	0.836	0.461	0.464	0.538	0.552	0.484	–
b_{res}	0.929	0.930	0.862	0.849	0.837	0.838	0.827	–
b_{saf}	0.999	0.999	0.601	0.609	0.619	0.639	0.469	–
b_{exp}	0.989	0.989	0.733	0.726	0.919	0.933	0.891	–

designs that would hardly be produced by a human designer. The single objective value overview of Table III shows that some objective functions are easier to optimize than others, and this appears to be consistent over the 2 map sizes: f_{res} is the most difficult, followed by f_{saf} and f_{exp}. While b_{saf} is near optimal for the MMO algorithms on small maps, it gets much harder to optimize for large maps. The b_{exp} and b_{res} values are very good for both map sizes. The average performance comparison in Figure 4 documents that for small maps, rES and NEA2 are almost on par concerning objective values, with NEA2 providing slightly more diverse maps. Novelty search and random search only differ in the average distance between selected maps. The large maps are more difficult to optimize, both objective values and distances are lower. Here, Novelty search performs similar in objective values, but worse in average distances. This behavior could be due to the high chances of infeasible offspring created from feasible parents, which causes the same feasible individuals (or small permutations thereof) to persist through several generations and be inserted multiple times in the novel archive, which in turns affects search.

Phase 2 Discussion: For the offline scenario with relatively long runs ($2{\cdot}10^4$ evaluations) considered here we can conclude that *it is* possible to reliably provide a set of good *and* diverse maps. The algorithm of choice is either rES or NEA 2 (both

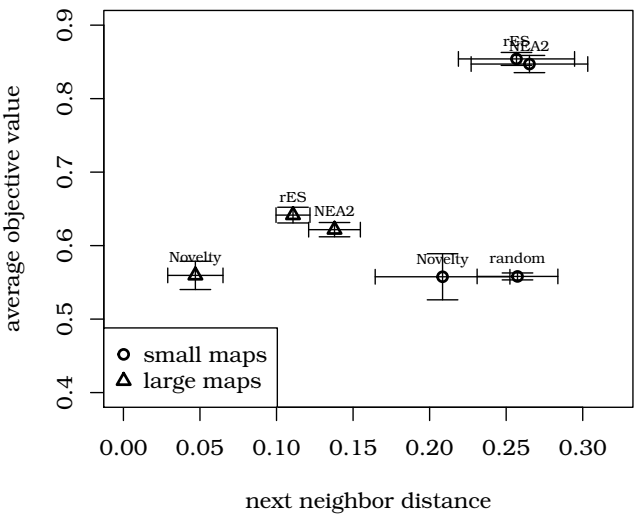


Fig. 4: Performance of the 4 algorithms in terms of average nearest neighbor VI-all distance and average objective values (21 runs), the error bars correspond to standard deviations.

Pareto optimal with respect to quality and diversity), depending on whether one wants to accept small quality losses for increased diversity. However, the degree of optimality differs largely for the different single objective criteria. When striving for a good average objective value, the algorithms deliver good performance for some of the criteria, whereas others are still far from optimal. Multi-objective optimization could be used to find out how these could be improved and what the cost with respect to the well optimized criteria would be.

VI. Overall Assessment

This paper has demonstrated that multimodal optimization algorithms such as NEA2 are able to create diverse map sketches of high quality, provided that there is a well-designed measure of diversity. In cases where the distance measure is ill-defined, however, niching is not effective and a standard restart EA can perform as well or even slightly better than NEA2. Finally, novelty search can create diverse individuals but their quality — even when searching for diversity in the dimensions of quality — is not as high as that of objective-driven evolutionary algorithms (either niching or non-niching).

The findings presented in this paper allow several conclusions to be drawn about the way diversity is perceived in game content, as well as how it affects algorithms which rely on it (such as novelty search or niching methods). However, we do not know how well the findings generalize to other game content domains, distance heuristics and parameters; this opens up several directions for future work.

The choice of map sketches for strategy games as the test domain allows for a succinct set of heuristics for calculating the quality of generated content. The objective functions used in Sentient Sketchbook have been tested extensively and have been, to a degree, verified by human game developers as important properties of strategy game levels. Additionally, their formulation allows for a smooth gradient which makes their optimization fast and straightforward for most stochastic search methods. While map sketches can also be used for other types of game levels such as first person shooters or platformers, the objective functions will need to be adapted to fit the level and may be less straightforward to optimize. Novelty search overall performed poorer than expected. The constrained space of map sketches might limit the performance of novelty search as it often fails to discover feasible individuals in large maps; more elaborate methods of constrained novelty search have been shown to perform better on the same problem, albeit with different search strategies [5]. Moreover the fact that novelty search uses a different scheme than the (1+20) of rES and NEA2 may also have contributed to its lower performance, considering that individuals were added to the novel archive on a per-generation basis. The generalizability of the algorithms, and of the findings regarding their performance, should be inspected with different parameters as well as in different domains either using map sketches, other types of game content which are less defined, or even in domains outside of games such as robot locomotion [20].

Regarding the measures of diversity, a small subset of the possible ways of describing the difference between content has been explored. As demonstrated, the diversity measure can affect the performance of both novelty search and niching algorithms quite substantially; additionally, the appearance of content generated can also be affected. Other alternatives to the existing diversity measures could include data from simulated gameplay. For instance, by having two artificial agents play against each other on a map sketch, gameplay features such as length of playtime or ratio of wins of one player over the other can be used to measure diversity (comparing a map biasing one player versus a map biasing the other). On the other hand, the visual impression distance has shown that some of its dimensions work better than others; exploring different combinations of existing dimensions of visual impression or introducing new ones may lead to even better results. While visual impression distance meshes well with perception theories in neuroscience [21] and cognitive psychology [9], evolutionary art [22], [23] can also provide measures for evaluating the difference between two 2D image representations. It is also possible to have the dimensions of visual impression distance automatically adapted, either during run-time or between runs (as a form of parameter tuning), based on the performance of the optimization algorithms. It is even conceivable that new visual features can be automatically defined via unsupervised learning of the algorithms' generated content; an early attempt of automatically discovered patterns which were used to define visual diversity for novelty search was presented in [24].

Finally, the experiments presented in this paper address the "offline" generation of game content [25], i.e. assuming no tight limitations on computational power or runtime. Novelty search has been shown to perform adequately in Sentient Sketchbook, where suggestions are generated on-the-fly with short evolutionary sprints of 10 generations [4]. Both rES and NEA2 algorithms, as described in this paper, require that one solution is optimized via local search before the algorithm moves on to the next solution. This means that for shorter evolutionary sprints these algorithms may underperform in terms of solution diversity: every local search is continued until it stagnates. For a real-time application such as Sentient Sketchbook, where solutions must be found within seconds,

novelty search might therefore still be the better alternative. Several solutions can potentially address this problem of scalability for rES and NEA2, such as optimizing each solution for a few generations in a round-robin fashion; future work should more thoroughly evaluate how such modifications to the core NEA2 method affect optimization, as well as how quickly solutions converge to "good" map sketches.

VII. Conclusion

This paper addressed the question of generating good *and* diverse content through a search-based approach. Several optimization algorithms and distance measures were systematically compared. It was found that the distance measure chosen has a crucial effect on diversity maintenance mechanisms, and in particular that variants of visual impression distance are good from an algorithmic standpoint; their appropriateness from a game design perspective, however, should be further verified by users of Sentient Sketchbook. It was also found that the NEA2 algorithm can balance quality and diversity well but only when using a good distance function, and that an evolution strategy with random restarts performed almost as well. Novelty search, which is specifically developed to maximize diversity of generated results, does not reach the same diversity and especially quality in the final artifacts as other methods. However, it is possible that the performance of novelty search could be improved with the right modifications, and it is also likely that it works relatively better with short runtimes.

From an optimization perspective, this is one of the first documented cases in which multimodal (niching) optimization algorithms actually provide an advantage for representations which are not numeric, and this opens up interesting avenues for more successful applications. We can even use niching algorithm performance to indirectly conclude on the "meaningfulness" of distance measures as demonstrated here. Not much can be said about a distance measure if niching fails, as the problem may be with the optimization algorithm. But if it succeeds, the measure is obviously useful, which is in turn a valuable insight into the problem domain.

References

[1] J. Togelius, N. Shaker, and M. J. Nelson, "Introduction," in *Procedural Content Generation in Games: A Textbook and an Overview of Current Research*, N. Shaker, J. Togelius, and M. J. Nelson, Eds. Springer, 2014.

[2] J. Togelius, M. Preuss, N. Beume, S. Wessing, J. Hagelbck, G. N. Yannakakis, and C. Grappiolo, "Controllable procedural map generation via multiobjective evolution," *Genetic Programming and Evolvable Machines*, 2013.

[3] A. Liapis, G. N. Yannakakis, and J. Togelius, "Towards a generic method of evaluating game levels," in *Proceedings of the AAAI Artificial Intelligence for Interactive Digital Entertainment Conference*, 2013.

[4] ——, "Sentient sketchbook: Computer-aided game level authoring," in *Proceedings of the 8th Conference on the Foundations of Digital Games*, 2013, pp. 213–220.

[5] ——, "Enhancements to constrained novelty search: Two-population novelty search for generating game content," in *Proceedings of Genetic and Evolutionary Computation Conference*, 2013, pp. 343–350.

[6] G. N. Yannakakis, A. Liapis, and C. Alexopoulos, "Mixed-initiative co-creativity," in *Proceedings of the 9th Conference on the Foundations of Digital Games*, 2014.

[7] S. O. Kimbrough, G. J. Koehler, M. Lu, and D. H. Wood, "On a feasible-infeasible two-population (fi-2pop) genetic algorithm for constrained optimization: Distance tracing and no free lunch," *European Journal of Operational Research*, vol. 190, no. 2, pp. 310–327, 2008.

[8] S. Björk and J. Holopainen, *Patterns in Game Design*. Charles River Media, 2004.

[9] R. Arnheim, *Art and visual perception: a psychology of the creative eye*, revised and expanded ed. University of California Press, 2004.

[10] S. Das, S. Maity, B.-Y. Qu, and P. N. Suganthan, "Real-parameter evolutionary multimodal optimization – a survey of the state-of-the-art," *Swarm and Evolutionary Computation*, vol. 1, no. 2, pp. 71–88, 2011.

[11] X. Li, A. Engelbrecht, and M. Epitropakis, "Benchmark functions for CEC'2013 special session and competition on niching methods for multimodal function optimization," RMIT University, Evolutionary Computation and Machine Learning Group, Australia, Tech. Rep., 2013.

[12] M. Preuss and S. Wessing, "Measuring multimodal optimization solution sets with a view to multiobjective techniques," in *EVOLVE – A Bridge between Probability, Set Oriented Numerics, and Evolutionary Computation IV*, M. Emmerich *et al.*, Eds. Springer, 2013, vol. 227, pp. 123–137.

[13] J. Lehman and K. O. Stanley, "Revising the evolutionary computation abstraction: Minimal criteria novelty search," in *Proceedings of the Genetic and Evolutionary Computation Conference*, 2010, pp. 103–110.

[14] ——, "Abandoning objectives: Evolution through the search for novelty alone," *Evolutionary Computation*, vol. 19, no. 2, pp. 189–223, 2011.

[15] N. Hansen, "The CMA Evolution Strategy: A Tutorial," version of June 28, 2011. [Online]. Available: http://www.lri.fr/~hansen/cmatutorial.pdf

[16] H.-G. Beyer and H.-P. Schwefel, "Evolution strategies – a comprehensive introduction," *Natural Computing*, vol. 1, no. 1, pp. 3–52, 2002.

[17] M. Preuss, "Improved topological niching for real-valued global optimization," in *Applications of Evolutionary Computation*, ser. Lecture Notes in Computer Science. Springer, 2012, vol. 7248, pp. 386–395.

[18] M. Preuss, L. Schönemann, and M. Emmerich, "Counteracting genetic drift and disruptive recombination in $(\mu +/, \lambda)$-EA on multimodal fitness landscapes," in *Proceedings of the 2005 conference on Genetic and evolutionary computation*, ser. GECCO '05. ACM, 2005, pp. 865–872.

[19] M. Preuss, P. Burelli, and G. N. Yannakakis, "Diversified virtual camera composition," in *Applications of Evolutionary Computation*, ser. Lecture Notes in Computer Science, C. Chio *et al.*, Eds. Springer, 2012, vol. 7248, pp. 265–274.

[20] J. Lehman and K. O. Stanley, "Improving evolvability through novelty search and self-adaptation," in *IEEE Congress on Evolutionary Computation*, 2011, pp. 2693–2700.

[21] V. S. Ramachandran and W. Hirstein, "The science of art: a neurological theory of aesthetic experience," *Journal of consciousness Studies*, vol. 6, pp. 15–51, 1999.

[22] S. Colton, "Evolving a library of artistic scene descriptors," in *Proceedings of the First International Conference on Evolutionary and Biologically Inspired Music, Sound, Art and Design*. Springer-Verlag, 2012, pp. 35–47.

[23] J. Correia, P. Machado, J. Romero, and A. Carballal, "Feature selection and novelty in computational aesthetics," in *Proceedings of the Second International Conference on Evolutionary and Biologically Inspired Music, Sound, Art and Design*, ser. EvoMUSART'13. Springer-Verlag, 2013, pp. 133–144.

[24] A. Liapis, H. P. Martínez, J. Togelius, and G. N. Yannakakis, "Transforming exploratory creativity with DeLeNoX," in *Proceedings of the Fourth International Conference on Computational Creativity*, 2013, pp. 56–63.

[25] J. Togelius, G. Yannakakis, K. Stanley, and C. Browne, "Search-based procedural content generation: A taxonomy and survey," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 3, no. 3, pp. 172–186, 2011.