Property-Based Testing for Visualization Development

M. Stegmaier1,∗, D. Engel2,∗, J. Olbrich1, T. Ropinski2, M. Tichy1

1Ulm University, Institute of Software Engineering and Programming Languages, Germany
2Ulm University, Institute of Media Informatics, Germany
∗equal contribution

Abstract
As the testing capabilities of current visualization software fail to cover a large space of rendering parameters, we propose to use property-based testing to automatically generate a large set of tests with different parameter sets. By comparing the resulting renderings for pairs of different parameters, we can verify certain effects to be expected in the rendering upon change of a specific parameter. This allows for testing visualization algorithms with a large coverage of rendering parameters. Our proposed approach can also be used in a test-driven manner, meaning the tests can be defined alongside the actual algorithm. Lastly, we show that by integrating the proposed concepts into the existing regression testing pipeline of Inviwo, we can execute the property-based testing process in a continuous integration setup. To demonstrate our approach, we describe use cases where property-based testing can help to find errors during visualization development.

CCS Concepts
• Human-centered computing → Visualization toolkits; • Software and its engineering → Software verification and validation;

1. Introduction
Testing plays a crucial role in ensuring software quality. When performed manually, it is labor-intensive and time-consuming. For this reason, it seems natural to automate testing as much as possible. Automated testing not only saves time, but also contributes to reproducible tests. In addition, the lower cost of automated testing compared to manual testing allows testing to be more thorough. In the context of test-driven development [Bec03], software tests are written in advance. A software requirement first gets converted to a test and only afterwards or in parallel the implementation of the requirement is developed.

While the automatic generation of tests is often straight-forward in the case of unit tests, it is significantly more difficult to test the interaction of different modules. This process is called feature interaction testing [PSK+10] (FIT). One of the difficulties that arise with FIT, is that through the many possibilities to combine modules for interaction, the resulting space of possible tests results in a combinatorial explosion. However, studies on debugging with feature interaction testing have shown that while testing the interaction of features in pairs already significantly increases the error detection rate, the increase from combining features in pairs to combining three or more features is insignificant [WK02,KWG04]. Thus, with pairwise coverage, a good balance between error detection rate and scalability can be obtained.

Data visualization techniques are used in a wide variety of research fields, such as medicine [KMM+18], biology [RE07], biochemistry [KKF+17] and physics [AAA+19]. As data visualization demands increase, and researchers as well as practitioners rely on visualizations to make critical decisions [KOJC13], it is of utmost importance to verify visualizations for their correctness. This includes software quality assurance for the development of the rendering algorithms underlying visualizations. While verification and rigorous testing is applied in a wide range of disciplines in computer science, the adoption of those processes has been rather slow in visualization [KS08]. This lack of rigorous software testing often prevents researchers from sharing their visualization prototypes, negatively impacting the reproducibility and practicality of their work. We believe the reason for this is that visualization algorithms are difficult to test due to their large set of parameters and their often hard to declare impact on the resulting rendering.

In this work, we focus specifically on the verification of the code implementations underlying visualization algorithms. Unfortunately, very little research has been conducted on appropriate testing concepts for visualization software. Current visualization software commonly makes use of unit tests and simple regression tests. We argue that those types of tests alone are insufficient to reliably verify the correctness of visualization implementations, as unit tests fail to capture errors occurring through the interaction of different software components. Current frameworks try to identify those integration errors using simple regression tests, where the resulting image of a visualization is compared to a reference image that was produced when the test was created. The problem with that
The remainder of this paper is structured as follows. We first discuss related work in Section 2, before introducing relevant prerequisites regarding Inviwo and property-based testing in Section 3. We detail our approach to adapting property-based testing for visualization development, on the example of Inviwo, in Section 4. In Section 5 we show possible use cases for our approach before we conclude in Section 6. Our contributions can be summarized as follows:

- We propose a concept to integrate property-based tests into visualization development.
- We improve upon the current state of testing implemented in visualization software, which only covers the parameter space of the visualization algorithms sparsely, as we sample larger portions of this space efficiently through automatic test generation.
- Our approach is compatible with Inviwo’s existing regression test system, which enables execution of tests whenever a code change is pushed to the public repository.

2. Related Work

In this section, we give an overview of prior research related to our approach, mostly from the field of verifiable visualization.

Verification is the process of assessing the correctness of a numerical solution to a given model [Roy05]. In the context of visualization, this usually means assessing the correctness of renderings against the underlying mathematical models. For example, a volume visualization algorithm can be considered correct if its results match the volume rendering integral [Max95]. The visualization community has started recognizing the need for verification in visualization already in the 90s [GU95]. Kirby and Silva [KS08] also advocate for a common framework for verification in visualization and introduce the term verifiable visualization. The verification process is typically divided into solution verification and code verification [KS08].

Solution verification focuses on evaluating the accuracy of the resulting visualization with regard to the continuous solution given by the volume rendering equation, and thus investigates the extent of errors introduced through numerical approximation. Note that many of the techniques used to verify visualization solutions are also well known in computer graphics [UWP06].

Etiene et al. [EJR∗13] investigate the correctness of a rendering with respect to the discretization of the volume rendering integral. They progressively refine step, grid and pixel sizes to analyze the convergence towards the continuous solution for direct volume rendering algorithms. In another work [ESN∗09], they also inspect the convergence rates of iso-surface features like surface normals, area and curvature for several iso-surface approaches using the method of manufactured solutions.

Zheng et al. [ZXM10] provide a framework to measure and correct errors in the reconstruction of projection-based data, like CT. More specifically they measure and bound the errors introduced through interpolation of off-grid data samples.

Code verification on the other hand focuses on detecting mistakes in the source code or the algorithmic implementation. Prior work on code verification in visualization is rather sparse, despite its wide-spread use in other disciplines of computer science. We are not aware of any published papers on the subject, however we can inspect existing visualization software to get an overview of the current state of code verification in visualization.
shrinking generated test cases still fail. Eventually, this results in a minimal a specified property, the tool simplifies the inputs as long as the generated test case identifies an error, i.e., the result does not satisfy this data to see if the specified properties hold true. After a generated test case, the shrinking process removes elements from the list and checks to see if the error still occurs. After the whole shrinking process, the reported list would be [3,3] which would erroneously get sorted to [3] by the faulty implementation. So by shrinking, we have obtained a minimal test case that already reveals the error in our implementation.

The best known implementation of property-based testing is QuickCheck [CH00] for the functional programming language Haskell. In QuickCheck, functions are used to define the expected properties. For existing types there are default generators that can generate random values. For user-defined types, as well as for existing types, QuickCheck allows for specifying custom generators. Following the success of QuickCheck for Haskell, its concepts have been adopted by libraries for several mainstream programming languages, such as C [GKS05], C++ [ee], and Java [PLS19].

In this paper, we describe how we adapt these concepts to enable property-based testing for visualization development. To avoid confusion in the terminology, we refer to the testable properties, in our case expected image changes, as effects and reserve the word property for Inviwo properties.

3.2. Inviwo

In this section we briefly introduce Inviwo [JSS*19] and its parts and concepts that are relevant to our implementation of property-based testing, as well as an overview of its existing testing pipeline. Inviwo is an extensible interactive visualization system that provides several usage abstraction levels. These abstraction levels range from developing low-level device-specific C++ code to modeling data-flow interactively at a high-level in Inviwo’s GUI. While the lower levels are used to extend the system with new capabilities through modules, the GUI can be used to create visualizations by using a network editor. The network resembles a directed acyclic graph with processors as nodes, representing functional units, and data-flow as edges, transferring data from a processor’s output to another processor’s input. A screenshot of the GUI showing such a network can be seen in Figure 3. The parameters of an algorithm which is implemented by a processor can be exposed using properties (see the window on the right in Figure 3). Processors can implement functions such as for instance data loading, filtering, rendering and displaying, allowing users to build custom visualization pipelines interactively using drag-and-drop. Properties of a processor can be all types of parameters associated to the processor’s function, ranging from file paths, over lighting parameters to transfer functions.

Testing in Inviwo is currently implemented through unit tests to validate individual modules and processors, as well as integration tests to validate a correct integration of multiple different proces-

---

Figure 2: The typical components of property-based testing.
Figure 3: The GUI to create tests with our implementation. We can define effects for all compatible properties from the previous processors of the network. In the case of Clip X Slices we expect the resulting number of pixels with background depth to be less or equal, when increasing the lower bound, and greater or equal when reducing the upper bound. The tested renderings of the clipped volume can be seen in Figure 6.

In common scenarios, Invivo allows for automatically generating regression tests from a given network, using the currently displayed canvas images as references to compare against in test runs, which are executed on a continuous integration server. Thus, code changes that break existing functionality can be easily spotted before being merged into the public repository. On top of that, Invivo allows for visual debugging inside the GUI by providing means to inspect common data types at any place in the network by hovering over an import or an outport. This allows for easy checking of intermediate results while developing a visualization.

One major drawback of Invivo’s current testing capabilities is that the aforementioned regression tests can only find errors occurring with the property values specified when creating the test. Therefore, the current approach will miss errors that only occur with different network parameters, i.e., with different values for the processors’ properties. This issue is addressed with our presented approach, which employs property-based testing to also cover this parameter space with tests.

4. Property-Based Testing in Visualization Development

In this section, we detail our approach to enable property-based testing in visualization development. We first describe the underlying ideas and concepts, and after that we explain how we implemented them.
of clipping planes, we determine the number of pixels with a given depth. The idea is that if there is a background and an object, all the pixels of the background have the same depth, and thus the number of pixels with that depth are the pixels that are not obscured by the object that is in the foreground. Changing the clipping planes will cause more fragments of the object to be discarded, and thus more pixels of the background to be visible.

In contrast to the testing methods currently existing in Inviwo, such as its unit tests, with property-based testing a user does not explicitly specify the properties of the processors for a test case, but specifies the expected effects for different values of a property. We let the user specify which properties should be tested and which effect in form of a comparison operator is expected by increasing the property. While property-based testing allows to define additional (randomized) unit tests, it also allows integration tests. In this context, property-based integration tests have the advantage that they can be defined from visual expectations alongside a visualization, using the same network, without the need to write code. As visualization development usually produces an image, it is natural to test visualization algorithms based on their effects on the resulting image.

In order for a property to be testable, there must be implementations for generating and comparing values or components of values. For the purpose of this paper, we provide default implementations for properties with a single integer or a single floating point value or a range of such. Upon testing, we generate up to nine values for each chosen testable property and combine them into sets of assignments. These nine values include numerical limits of the data type as well as uniformly chosen random samples within those limits.

In a network there can be several processors, each of which has several properties. Since several random values are generated for each of these properties, a huge number of combinations is possible. By restricting the combinations to pairs, this number is significantly reduced. For each chosen property we generate tests, such that each possible pair of values for different properties is in some test. That means the generated set of tests satisfies pairwise coverage. Two tests are comparable, if and only if the expected effects of the individual properties are non-contradictory, that is, there is a combined expected effect that implies each individual effect. For example, in the case of our clipping example, if there is a property that controls the opacity of the rendered object. A value of 0 would make the object completely invisible and thus make the clipping have no effect anymore.

For the generation of pairwise coverage, we use an algorithm for covering array construction, which in our case constructs an array containing test cases, so that each pair of random values is covered. At first, we constructed our covering array using the discrete SLJ strategy from [SC19]. Typically, the goal of algorithms for constructing covering arrays is to keep them as small as possible, and this algorithm achieves that quite well, but it turned out that these minimal sets unfortunately also contained only very few comparable pairs of tests.

Since for our purposes, the number of comparable pairs of tests should be as large as possible (for a fixed number of tests), we developed our own greedy algorithm which not only aims to keep the total number of tests low, but simultaneously also aims to maximize the number of comparable pairs of tests. Let \( A_1, A_2, \ldots, A_t \) be the sets of possible assignments for the tested properties, and let \( A^t = A_1 \cup \ldots \cup A_t \), where \( \bot \) represents an undefined assignment. We call the elements of \( A = A_1 \times \cdots \times A_t \) (complete) tests and the elements of \( A^t = A_1^t \times \cdots \times A_t^t \) partial tests, and we say a (partial) test \( t = (a_1, \ldots, a_t) \) implies another (partial) test \( b = (b_1, \ldots, b_t) \) if and only if \( b_i \in \{ \bot, a_i \} \) for all \( i \in \{1, \ldots, t\} \), and \( a_i \) and \( b_i \) are disjoint, if and only if \( (a_i \neq \bot) \land (b_i \neq \bot) \implies a_i = b_i \) for all \( i \). We say \( c = (c_1, \ldots, c_t) \) arises from extending \( a \) with \( b \), where \( c_i = a_i \) if \( a_i \neq \bot \) and \( c_i = b_i \) otherwise.

Note that if \( a \) and \( b \) are disjoint, \( c \) implies both \( a \) and \( b \). Our general approach is shown in Algorithm 1. When successively constructing our covering array \( C \subseteq A \), we first generate all possible interactions \( I \),

\[
I = \{(a_1, a_2, \ldots, a_t) \in A^t : \forall i \in \{1, \ldots, t\} \}
\]

between two properties, and while there are still unused (pairwise) interactions (i.e., interactions in \( I \) that are not implied by some already generated test), we construct a new test \( v \) which initially is equal to one randomly selected unused interaction. While \( v \) is not complete, we extend it with either a disjoint and not implied unused interaction or a (not necessarily disjoint) previously created test, such that the resulting partial test maximizes the sum of weights of all comparable previously created tests. Here, the weight of a test is equal to the current number of created tests to which it is not comparable. This is intended to increase the minimum number of other generated tests, that a test is comparable to, such that there are as few tests as possible for which there are no other comparable tests. Also, the weight of unused interactions is multiplied by a given constant in order to “encourage” using multiple uncovered interactions in one test and thereby reduce the total number of tests.

Note that for any incomplete test there is always a (possibly already used) disjoint interaction, so our algorithm is guaranteed to terminate and output a 2-covering array.

Since there can be many processors in a network, each of which can have multiple properties, there can be an unmanageable amount of randomizable values. It would be difficult to determine which combination of parameters actually caused an identified error. To solve this, we adapted the concept of shrinking, as introduced in Section 3.1. When an error is identified, we reset the randomized properties one by one to their original value and test to see if an error still occurs. In the end, we will have a minimal set of randomized properties that still causes an error.

The concepts described above can easily be adapted to any other rendering pipeline that exposes its properties in a way that allows them to be varied programmatically. In the following we describe our exemplary implementation of these concepts as an Inviwo processor.

### 4.2. Implementation

We have implemented property-based testing in Inviwo by realizing a processor called PropertyAnalyzer. This processor receives the image produced by the network that we want to verify, and gathers the modifiable properties from all processors involved in producing that image. These gathered properties can then be used to define tests through the properties of the PropertyAnalyzer processor as...
input : Sets of possible assignments \( A_1, \ldots, A_r \), weight multiplier \( m \)
output: A 2-covering array 
\[ C \leftarrow \emptyset; \]
while there is uncovered interaction \( u \) do
\[ v \leftarrow u; \]
while \( v \notin A \) do
\[ w_{\text{tmp}} \leftarrow \infty; \]
for uncovered interaction \( u' \) do
\[ \text{if } \text{disjoint}(v,u') \land \text{implies}(v,u') \land m \cdot \]
\[ wS(C, \text{extend}(v,u')) > w_{\text{tmp}} \text{ then} \]
\[ \text{tmp} \leftarrow \text{extend}(v,u'); \]
\[ w_{\text{tmp}} \leftarrow m \cdot wS(C,\text{tmp}); \]
end
for \( c \in C \) do
\[ \text{if } wS(C, \text{extend}(v,c)) > w_{\text{tmp}} \text{ then} \]
\[ \text{tmp} \leftarrow \text{extend}(v,c); \]
\[ w_{\text{tmp}} \leftarrow wS(C,\text{tmp}); \]
end
\[ v \leftarrow \text{tmp}; \]
end
\[ C \leftarrow C \cup \{v\}; \]
end
\[ \text{return } C; \]

Algorithm 1: Our algorithm for the generation of the covering array. With \( \text{weight}(C,a) = \{c \in C : \text{not comparable}(c,a)\} \)
the number of tests in \( C \) that are not comparable to \( a \) and \( wS(C,a) \) being the sum of the weights of all tests in \( C \) that are comparable to \( a \), i.e. \( wS(C,a) = \sum \{\text{weight}(c) : c \in C, \text{comparable}(c,a)\} \)

Figure 5: Our PropertyAnalyzer processor generates a textual description of the clipping test.

When there is a pair of tests for which the combined expected effect contradicts the observed effect, we have found an error, that is, a case where the network does not behave as specified by the user. In order for the user to evaluate the errors, we generate a report document (in HTML) containing all property values as well as the
In this section, we illustrate the effectiveness of property-based tests for developing visualization techniques. Therefore, we describe several use cases where property-based testing can help finding implementation errors. The clipping use case described throughout this paper, is presented in Section 4. Note, that after integrating the presented property-based test use cases into Inviwo’s regression test system, they can be verified on the continuous integration server after every code change pushed to the public repository.

5.1. Lighting Parameters

Another excellent use case for property-based testing is varying lighting parameters to find errors in the rendering algorithm. Although not all parameters of a rendering technique have an easily testable effect on the resulting image, we can capture some common effects that should apply to all rendering techniques. In general, increasing or decreasing light intensity results in a monotonic increase or decrease in image brightness respectively.

A specific example with the Phong shading parameters - ambient, diffuse and specular lighting - is illustrated in Figure 7. A correctly implemented rendering algorithm using this model should always produce an image with the same or higher/lower pixel-wise brightness when increasing/decreasing these parameters. Note that this effect is true for differences in one of those components at a time, but not necessarily when comparing two renders where multiple of the components vary in different directions, i.e., one with low ambient and high diffuse and another vice versa.

In order to test for that with our implementation, one would first need to implement pixel-wise comparison. Given that, a test can be designed using an individual PropertyAnalyzer processor for each of the lighting parameters, ambient, diffuse and specular, respectively. Each PropertyAnalyzer verifies the effect for its lighting property, while setting the others to NOT_COMPARABLE, in order to avoid tests with parameter sets that vary in multiple lighting properties simultaneously. Setting the other components to NOT_COMPARABLE will still generate tests with different values for those, but avoids the aforementioned issue by excluding comparisons where multiple components are changed at the same time.
Normal
Low ambient
High diffuse
High specular

Figure 7: Increasing or decreasing the brightness of lights and materials should increase or decrease the brightness in the resulting image, respectively.

(a) normal  (b) low ambient
(c) high diffuse  (d) high specular

Transfer function used for the transparent rendering (left). The opacity of the selected control point is increased to get the opaque rendering (right).

Figure 8: Increasing or decreasing the opacity of the transfer function should result in an increase or decrease in alpha, respectively.

5.2. Compositing

Manually finding errors in the integration of light along a ray can be very difficult, especially when transparency effects are involved. While it is impossible to test all aspects of this complex task, there are again common assumptions that we can make, that should apply to all of our integration approaches. One of those assumptions is that, when increasing the opacity of the material, as defined by a transfer function, the resulting image should monotonically increase in opacity. This effect is illustrated in Figure 8.

In order to test that, we can manipulate the transfer function. In Inviwo transfer functions can be defined as piece-wise linear functions using control points in a widget, as illustrated in Figure 8c. After implementing an appropriate input generator for the transfer function, the PropertyAnalyzer can generate assignments with two transfer functions that vary in the opacity of a single control point. Each pair of renderings is then checked for our assumption, that a larger opacity in the transfer function should lead to an equal or higher pixel-wise alpha. Note, that this assumption only holds, when varying a single control point at a time, that means all other control points should be set to NOT_COMPARABLE.

5.3. Limitations

However we also found limitations of our approach. When relying only on the produced rendering to determine the correctness of an effect, we can miss implementation errors complying with the desired effects. In the clipping example, our approach would be unable to detect an error if two axes are switched, because clipping a different axis still has the same effect on the aggregated image metrics. Our approach should be used complementary to existing unit and regression testing strategies, in order to increase their parameter coverage, but it should not be seen as a replacement for other types of tests.

6. Conclusion

In this work, we propose the integration of property-based testing into visualization development. We argue that property-based testing can be used to vastly increase the coverage of different rendering parameters in the testing procedure. This allows us to find implementation errors quicker and increases the confidence in the correctness of our implementations. In fact, our proposed property-based testing implementation can even be used to create visualizations in a test-driven manner, by defining the tests alongside the actual algorithm. With this new testing strategy, we offer an easy way for visualization researchers and developers to create meaningful tests automatically from within a GUI, just by defining their expectations for parameter changes. By making the testing process easier and more expressive, we expect to further close the gap between research and software development in the field of visualization.

In the future, we plan to further extend the capabilities of our implementation through new types of effects, as well as more convenient ways to generate multiple, very similar tests on the fly, like the Phong lighting example from Section 5.1. Another interesting direction to follow is the development of an interface, that allows users to interactively choose additional reference images for regression tests, based on property-based testing parameter generation, as an easy way to cover a larger parameter space through regression tests (using image comparison) directly.

Acknowledgments

This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG) under grant 391107954 (Inviwo). The proposed concepts have been realized using the Inviwo visualization framework (https://inviwo.org).