An Empirical Evaluation of Automated Black-Box Testing Techniques for Crashing GUIs

Cristiano Bertolini Glaucia Peres Marcelo d'Amorim Alexandre Mota

Center of Informatics, Federal University of Pernambuco
P.O. Box 7851, 50732-970, Recife-PE, Brazil
E-mail: {cbertolini,gbp,damorim,acm}@cin.ufpe.br

Abstract

This paper reports an empirical evaluation of four black-box testing techniques for crashing programs through their GUI interface: SH, AF, DH, and BxT. The techniques vary in their level of automation and the results they offer. The experiments we conducted quantify execution time and the capability of finding a crash for each technique on 8 different cellular phone configurations with historical (real) errors. The results show that AF and BxT offered better precision (i.e., the fraction of runs that end in a crash out of the total number of runs) than SH and DH (AF and BxT found crashes in all 8 configurations), and BxT crashes the application the fastest more often (5 out of 8 cases). The experiments reveal that the selection of the random seed to AF and BxT results in a high variance of execution time (i.e., the time the technique takes to either crash the application or timeout in 40h): the mean (across 8 phone configurations) of the standard deviation of execution times (for 10 runs per each phone configuration) is 7.79h for AF and 5.21h for BxT. Despite this fact, AF and BxT could crash the application consistently: the mean of the precision (fraction of the 10 runs that results in a crash) is 74% for AF and 69% for BxT.

1. Introduction

Despite the technological advances in languages and tools to support system's development, programmers still deliver software with lots of errors. Several techniques have been proposed to this end – to improve software reliability. Testing is one of them. In fact, software testing is the dominant approach in industry to assure software quality. Testing is not cheap though. Santhanam and Hailpern [13] reported that from 50% to 75% of the total cost of a project involves testing and debugging.

Black-box box testing is the activity of testing without knowledge of the program organization [5]. It consists of exercising the interface of a component (typically the entire system) to find errors. Any part of the system providing a public interface is amenable to black-box testing. White-box testing, in contrast, requires knowledge of the program structure. For example, with white-box testing an engineer may construct a test with the specific goal to exercise a program path. Black-box and white-box testing are recognized as complementary techniques [12]. However, it is important to note that the testing team is not able to apply white-box testing when the application owner prohibits the delivery of source code, say for confidentiality reasons.

One important interface of an interactive system is the graphical user-interface (GUI). A GUI test consists of (i) a sequence of GUI commands and (ii) a test oracle to check whether the execution of this sequence produces the expected result [11]. In our context, the oracle is conceptually a boolean function that checks whether the application can still make progress through the GUI.

This paper focuses on black-box testing of GUIs. More specifically, this paper proposes and evaluates several techniques to *automate* the generation of GUI tests. We investigate techniques whose *goal* is crashing the system with the *automated generation and execution of GUI tests*. Automated test generation is important for two main reasons: (a) manual tests can become obsolete with the evolution of an application, and (b) the quality of manual tests depends on the level of completeness of requirements. We are particularly interested on addressing the second problem. For example, manually-written (system) tests (derived from requirements) may succeed in covering common user interactions but fail to cover corner case scenarios that can lead to a crash [4, 10].

Automation of black-box testing can be challenging: unguided search may be ineffective for too large state spaces [8]. For GUI testing, in particular, the size of the state space reachable (from the GUI) is typically intractable [6, 15]. To alleviate the state space explosion problem many model-checking techniques need to access the state to per-

form space reductions [7, 9, 14]. Unfortunately, such optimizations are not possible for black-box testing in general.

Our *context* is that of applications providing limited or no access to its internal state. This paper describes four black-box testing techniques for finding program crashes on GUIs. All techniques attempt to *explore the state space* of the application (i.e., to *stress* the application with automated interactions) with the goal of finding a state that fails the test oracle. The techniques we propose provide distinct tradeoffs between their capability of finding crashes and the level of automation they offer. The main focus of this paper is on the evaluation of these techniques on Motorola cellular phones. The list below highlights the main contributions:

- The proposal of four techniques for GUI testing.
- An empirical evaluation of the techniques using Motorola cellular phones.

We provide next a brief overview about the techniques and the experimental results we obtained. Section 2 discusses the techniques in more detail and Section 3 the experimental evaluation.

1.1. Summary of the techniques

We next summarize the techniques we propose for crashing applications from the GUI: (i) Scenario Hunting (SH), (ii) Atoms Framework (AF), (iii) Driven Hopper (DH), and (iv) Behavior eXplorer Tool (BxT). It is important to note that we do not distinguish between "test" and "test sequence" as the notion of correctness we use is universal: the program should not crash. As such, the oracle is not part of one test artifact.

SH takes a manually-written test suite as input, generates a fixed number of random permutations of tests in this suite, and finally executes each test and monitors for a crash. SH is perhaps the simpler technique we discuss. Its merit relies mostly on the user experience for selecting tests in the input suite. It serves as a baseline to compare more automated techniques.

AF also takes as input a manually-written test suite. AF differs from SH in two important ways: (i) the granularity of the tests it uses from the input suite, and (ii) how it builds input data. As for test granularity, one AF test corresponds to a small fragment of a SH test. For that reason we use the term *atom* in reference to one AF test. As for data generation, AF enables a test to share data: one test can consume data another test produces. In contrast, one SH test can only consume data it generates.

Illustrative example. Figure 1 shows a fragment of one test (sequence) that SH uses in the evaluation we conducted on cellular phones (see Section 3). This test is written in Java and runs on a regular PC connected to a phone. This sequence consists of using one phone to (i) capture an audio message, (ii) play back that message, and (iii) delete that same message. To enable AF the engineer needs to divide this test in smaller self-contained fragments. For this example, *the engineer* uses the log instruction to identify these fragments; three atoms in this case. Note that one AF test (atom) may require parameters in result of this method extraction. For example, the second atom (for listening the audio message) will require a *multimediaFile* object. This is key to AF as it enables one atom to exercise different inputs. Section 2.2 details AF, including how it combine atoms to build larger sequences. SH and AF build on the user experience to find crashes. The following techniques, in contrast, require less user-input.

```
log("Capture an audio message.");
navigationTk.launchApp(PhoneApplication.get("VOICE_RECORDS"));
multimediaTk.captureVoiceNoteFromVoiceRecord(30);
multimediaFile voiceRecord =
multimediaTk.storeMultimediaFileAs(MultimediaItem.get("STORE_ONLY"));
log("Listen to an audio message.");
navigationTk.goTo(PhoneApplication.get("VOICE_NOTES"));
multimediaGoTo.get("ALL_VOICE_NOTES"));
multimediaTk.scrollToAndSelectMultimediaFile(voiceRecord);
log("Delete an audio message.");
phoneTk.returnToPreviousScreen();
multimediaTk.deleteFile(voiceRecord,true);
```

Figure 1. A test sequence for multimedia.

DH drives the application to a particular screen and keeps pressing random keys (some of which can change the current screen) for some (configured) time until it finds an error or crashes the application. DH requires tests to drive the phone to an initial screen. Such tests use the instruction `goto()` from Figure 1 for this setup.

BxT attempts to make a more systematic selection of inputs than DH: it recognizes which controls are available at a screen and selects inputs according to these controls. For instance, BxT can send scroll down and up events when it recognizes a scroll bar control in the current screen. In a screen that contains only two buttons, say "OK" and "Cancel", DH may press several keys before it hits the ones for "OK" and "Cancel". BxT, differently, makes a random selection between one of these two options. Note, however, that BxT has more stringent observability and controllability requirements [5,12] than DH: it requires a library providing support for recognizing screen components and sending specific events to them.

1.2. Summary of results

We evaluate the techniques on cellular phones with historical (real) errors. We run each technique for 10 times with different seeds over each configuration. Our empirical

Algorithm 1: genList pseudo code

```
1  genList(Set⟨Test⟩ suite, int numRept, int seed): List⟨Test⟩
2  begin genList
3      List⟨Test⟩result = [ ];
4      for i = 1..numRept do result = result.add(shuffle(suite, seed));
5      return result;
6  end
```

results demonstrate that AF and BxT together outperformed SH and DH with respect to time and also to the number of crashes reported.

We use the term precision to denote the fraction of runs that ends in a crash out of the total number of runs. The precision for AF and BxT was 74% and 69% respectively. This result indicates that a more automated technique (BxT) performed nearly the same w.r.t. precision as one using user-provided test suites as input (AF). Section 3 details this experiment and others we conducted to better understand how each technique performs.

2. Techniques

This section describes four testing techniques this paper proposes for crashing applications through their GUIs.

Note on oracle. To simplify discussion, we assume crashes are unexpected situations which the system can not continue its normal execution. That allows the algorithms to represent the oracle with the external function *isCrash()*. We do not discuss this function here. Note that this paper does not propose test oracles. The user needs to provide the oracle appropriate for detecting crashes.

Note on user-provided test suite. SH, AF, and DH build on existing manually written tests. But DH tests simply perform a jump to one GUI screen.

Note on GUI library. DH and BxT build on operations (provided by some library) that enables some read and write access to the GUI components. Sections 2.3 and 2.4 highlight the operations DH and BxT use to clarify how they can be used in different contexts. It is important to mention that some systems provide rich support for testing, i.e., specific interfaces for reading (i.e., observing behavior) and writing to the state (i.e., controlling the application). For example, the cellular phone platforms Symbian [3] and Linux/Java [2] provide infrastructure to the tester implement monitors that can inspect the memory for safety problems such as buffer overflows and memory leaks.

2.1 SH

Algorithm 1 generates a random list of tests from a set of user-provided tests. Each iteration of the loop at line 4 generates one permutation of the input set of tests. Effectively, it provides as result a list that includes *numRept* permutations of *suite*. Section 3.4.2 elaborates on a variation of this algorithm.

Algorithm 2: SH pseudo code

```
1  main(Set⟨Test⟩ suite, int numRept, int seed, int timeout): bool
2  begin main
3      List⟨Test⟩ testList = genList(suite, numRept, seed);
4      foreach test in testList do
5          test.run();
6          if isCrash() then return true;
7          if isTimeout(timeout) then return false;
8      endfch
9      return false;
10 end
```

Algorithm 3: AF pseudo code

```
1  main(Set⟨Test⟩ suite, int numRept, int seed1, int seed2, int timeout):
   bool
2  begin main
3      Map⟨String, List⟨Object⟩⟩ dataMap = loadDataMap();
4      Set⟨Atoms⟩ atomSet = ∅;
5      foreach test in suite do atomSet = atomSet ∪ test.atoms();
6      List⟨Test⟩ testList = genList(atomSet, numRept, seed1);
7      foreach test in testList do
8          dataMap = test.run(dataMap, seed2);
9          if isCrash() then return true;
10         if isTimeout(timeout) then return false;
11     endfch
12     return false;
13 end
```

Algorithm 2 shows the pseudo code for SH. Function *main* assigns the result of the call to *genList* to variable *testList*. Each iteration of the loop at line 7 executes one test from this list, checks for a crash, and checks for timeout. Execution either terminates reporting a crash (line 6), or reporting a timeout (line 7). Also, SH can run without a crash or timeout that means all tests were executed and no crash was found (line 9).

Note on suite selection. The selection of the input test suite (*suite*) is decisive for the final result. Conceptually, the number of tests in the suite affects positively the chances one important part of the application is exercised (i.e., contains the defect) and negatively the exhaustion to which this part is exercised (i.e., may fail to activate the defect).

2.2. AF

Algorithm 3 shows the pseudo code for AF. The map *dataMap* that function *main* declares provides data input for the execution of tests. This map associates a list of objects to each input category (the map key). One atom is a parametric test that consists of a user-defined fragment from a user-defined test. The execution of one atom (test) may read

Algorithm 4: DH pseudo code

```
1 main(List⟨Test⟩ screens, int seed, int timeout1, int timeout2): bool;
2 begin main
3   while !isTimeout(timeout2) do
4     Test screen = listOfScreens.pickOne(seed);
5     screen.run();  /*goto random screen*/
6     pressRandomKeys(seed, timeout1);                    ⇐
7     if isCrash() then return true;
8   endw
9 end
```

from or update the data map.

AF stores in variable *atomSet* a set of atoms derived from the tests in *suite*. The variable *testList* stores the list of atoms resulted from the call to *genList*. Similar to SH each iteration of the loop at line 7 executes one test from *testList*, checks for a crash, and checks for timeout. However, different from AF a test takes as input the data map *dataMap* and the seed *seed2* and produces a new data map, possibly extending the input map with new inputs. The seed allows a test run to randomly choose one input from a list of objects for a specific category.

In summary, AF differs from SH in two important ways: (i) test granularity (atoms are fragments of SH tests), and (ii) data generation (the execution of one test provides inputs to parametric tests).

2.3 DH

Algorithm 4 shows the pseudo code for DH. The algorithm repeats the following sequence of steps until it either timeouts or finds a crash: (i) selects one screen, (ii) drives the application to that screen, (iii) sends random events (key presses) to the GUI for a while, and (iv) checks for a crash. The loop at line 3 repeats this sequence of steps.

The inputs to DH are a sequence of manually written tests – *screens*, a random seed that the event generator uses – *seed*, a bound on execution time for sending random events (key presses) in one iteration – *timeout1*, and a bound on total execution time – *timeout2*.

Note that DH does *not* generate inputs (i.e., GUI events) according to the components active on the current screen. It simply generates control events *randomly* within one important region of the application. Also important to note is that the algorithm uses a library to send events to the GUI. For DH it sends *general* key pressed events – which perform well for the domain of cellular applications (see Section 3.3). Line 6 highlights the use of one library function for sending key pressed events to the application. One needs to provide such a function to enable DH.

Algorithm 5: BxT pseudo code

```
1  main(int seed, int numRept, int timeout): bool;
2  begin main
3    Set⟨Test⟩ screenSet = ∅;
4    if driven() then
       /*random set of goto-screen tests*/
5      screenSet = {tc_1, tc_2, ..., tc_n};
6    else
7      screenSet = {initialScreen()};
8    endif
9    while !isTimeout(timeout) do
10     screenSet.pickOne(seed).run();
11     for i=1 to numRept do
12       Event ev = enabledEvents().pickOne(seed);      ⇐
         /*sends message to the GUI*/
13       ev.genInputs(seed).run();                      ⇐
14       if isCrash() then return true;
15     endfor
16   endw
17   return false;
18 end
```

2.4. BxT

Algorithm 5 shows the pseudo code for BxT. The main difference from DH is the way it generates events. Conceptually, the algorithm contains two main parts. The first decides which screen to focus. The second part stresses the application from a selected screen. This part performs the following steps for a fixed number of times or until it crashes the application: (i) identifies enabled events on the *current* screen (i.e., the events that active components can process), (ii) selects randomly one of such events, (iii) generates data for this event and sends the event to the GUI, and (iv) checks for a crash. The code fragment in the line range 12-14 corresponds to this sequence.

The inputs to BxT are a seed used for generating sequence (i.e., choosing the event) and data (i.e., generating input to the event) – *seed*, an integer denoting the number of iterations on the second part of the algorithm – *numRept*, and a bound on the total time for testing – *timeout*.

Line 10 makes a random choice of which screen in the set *screenSet* execution should stress. Note that the code fragment in the line range 4-8 initializes this variable. The external boolean function *driven()* indicates whether or not execution should perform jumps across different screens (in a similar fashion to DH). We call driven BxT, or simply D-BxT, this variation of BxT. For D-BxT, execution initializes the variable *screenSet* with a fixed set of screens. (This is how we used D-BxT in our experiments. For clarity, we showed the initialization of *screenSet* within the algorithm.)

Lines 12 and 13 highlight the uses of a library to identify which events are enabled and to send the event to the GUI.

Config.	SH		AF		DH		BxT		D-BxT	
-	CID	time	CID	time	CID	time	CID	time	CID	time
A	-	40.0	6	**6.5**	1	21.2	6	14.6	6	33.8
B	-	40.0	4	33.6	-	40.0	-	40.0	6	**6.2**
C	-	40.0	4	36.5	-	40.0	-	40.0	12	**14.8**
D	2	4.3	7	5.0	8	3.3	7	**1.2**	8	4.4
E	3	3.9	1	3.6	3	7.0	5	**0.5**	5	2.7
F	-	40.0	1	**4.0**	6	5.7	-	40.0	11	**4.0**
G	4	3.1	6	11.2	6	22.7	6	8.0	1	**1.7**
H	5	**2.3**	5	2.8	5	21.7	4	12	10	7.9
avg.	50%	21.7	**100%**	12.9	75%	20.2	62.5%	19.5	**100%**	9.4

Table 1. Time (in hours) and Fault revealed by SH, AF, DH and D-BxT per phone configuration.

3. Evaluation

This section provides details on the empirical evaluation of the techniques using Motorola cellular phones. We conducted 3 sets of experiments. One for comparing the four techniques w.r.t. their capability to crash phones with known bugs (Section 3.3). For this experiment we use SH and DH as baselines. The experimental results indicate that AF and D-BxT outperform SH and DH. Section 3.4 discusses the impact of randomization (for data and sequence generation) on AF and Section 3.5 discusses the impact of randomization on BxT. Section 3.6 discusses whether a more uniform exploration of the screens correlates with the time to find a bug on D-BxT.

3.1. Characterization of subjects

We characterize each subject with (i) the phone model (i.e., a list of external and internal phone features to identify a set of similar phones functions), (ii) the hardware version, (iii) the software version (i.e., the build for the operating system and its applications), and (iv) the flex bit (FB) version. The flex bit configuration allows the user to dynamically configure the phone prior. Example of such configurations includes enabling the phone to send and receive bluetooth signals, and setting the phone to debug mode.

Config.	Model	Hard.	Soft.	FB
A	M1	H3	S1	F1
B	M1	H4	S2	F2
C	M1	H4	S3	F3
D	M2	H2	S4	F4
E	M2	H1	S5	F5
F	M3	H5	S6	F6
G	M3	H6	S7	F7
H	M2	H2	S8	F8

Table 2. Characterization of experimental subjects.

Table 2 shows the subjects we used in our experiments. Column "Config." introduces a unique identifier to distinguish each combination of model, hardware, software and flex bit. The other columns show each of these attributes. The identifiers we use in this table correspond to real identifiers, but they are masked for confidentiality reasons. Note that some configurations share the same model or hardware, but the software and flex bit vary. The selection of these configurations was driven by the availability of equipment where past errors have been detected.

3.2. Failures

The oracle does not operate on the GUI. It is a general Motorola proprietary program that monitors the phone memory for bad states.

The oracle detects 12 distinct kinds of crashes across all experiments. In the following, we distinguish them using crash identifiers (CIDs) from 1 to 12. Each identifier denotes a different undesirable scenario of the application that the oracle is able to capture. For example, CID=1 is a general report to denote that the system makes no progress but the oracle is unable to ascertain the reason, CID=2 means that an issue with the hardware interface (e.g., it is not possible to allocate memory) prevents the application from making progress, CID=6 denotes a programming error like divide by zero, etc. Important to note is that the oracle reports only the crash event; it does not inform the reason for the crash (as debuggers do). In result, it may happen that distinct techniques report different manifestations (CIDs) of the same defect.

3.3. Comparison of techniques

This section describes the experiment we conducted to compare the techniques.

Setup. The goal of this experiment is to compare the effectiveness of AF, BxT and D-BxT with that of SH and DH for crashing cellular phones with historical defects. We

used 8 different phone configurations for which SH found 4 crashes and DH found 6 crashes. Neither SH nor DH crashed 2 of the eight configurations. For each configuration we ran once each technique until execution runs out of time (timeout=40h) or finds a crash. The execution of SH and DH confirmed the crashes documented in the bug report database. For AF, BxT and D-BxT, we fixed the random seed across different configuration runs.

The *atoms* that AF uses derives from the tests SH used – we did not include any atoms original from a different set of tests. This helps us to compare SH and AF. BxT and D-BxT explores the state space similarly to DH – exercise a random sequence of events on the GUI for 30s from one arbitrary GUI screen. To achieve this we align the setting of the paramaters *timeout1* in Algorithm 4 and *stepSize* in Algorithm 5. This similarity helps us to compare DH, BxT and D-BxT.

Results. Table 1 shows a summary of the results obtained in the experiments. Column "Config." shows the identifier for one subject configuration, column "CID" shows the identifier of the crash, and column "time" shows the execution time for each experiment. Recall that one experiment either timeouts or finds a crash. Line "avg." reports the averages of each column. For column CID, it shows the fraction of experiments that revealed a crash. For column time, it shows the arithmetic mean of the elapsed time.

We list below our key observations:

- Only AF and D-BxT can find a crash for all eight experiments with a timeout of 40h. As such, note that only AF and D-BxT can find the crashes reported in experiments B and C.

- D-BxT can find a crash faster more often than the other techniques. Note (from the highlighted cells) that D--BxT outperforms the other techniques in 4 out of 8 cases.

- SH can find a crash in only 50% of the cases. But when it finds, it outperforms AF, except in Config. E where the difference is of only 0.3h (that is, 18min).

- For each experiment where DH finds a crash, one of the other three techniques can find it and find it faster. In particular, AF + D-BxT find all errors that DH finds and faster, except in Config. D and E.

- The variation of time that each technique reports is very high. For example, between D-BxT and AF the difference in time for crash for experiment A is +27.3h (i.e., D-BxT takes 27.3h more to find the crash), -27.4h for experiment B, -21.7h for C, -9.5h for G, and +5.1h for experiment H.

- The variation of the CID reported is also high. Note that no experiment reports the same CID for all techniques. For example, in experiment H, SH, AF, and DH report CID=5, while BxT reports CID=4 and D-BxT reports CID=10.

3.4. Impact of Randomization on AF

This section discusses two experiments we conducted to evaluate the impact of randomization on AF. The first experiment measures the impact of the random data on the effectiveness of AF. For this, we vary the value of parameter *seed2* used in line 8 of Algorithm 3. The second experiment evaluates the impact of a random selection of the list of *atoms* when compared to the list computed from SH tests used in the experiments Section 3.3 reports.

Note on distribution representation. We use boxplot notation to illustrate a data distribution. The lower and upper hinges of one box indicate respectively the upper bounds of the first and third quartiles of the distribution, the line across the box defines the second quartile (i.e., median value). The lines below and above the box limit the first and fourth quartiles. Small circles outside the hinges correspond to outliers. The symbol $\bar{x}$ denotes the mean value, the symbol σ denotes the standard deviation – an average for the dispersion of data points from the mean value, and the symbol $\hat{x}$ denotes the median value.

3.4.1 Random data

This section describes the experiment we conducted to evaluate the impact that the use of different random data has in the effectiveness of AF.

Setup. In this experiment, we run AF for 10 times, on each configuration, varying the value of the parameter *seed2* used in Algorithm 3. We use the same sequence in all executions of a configuration (i.e., we fix the values of *seed1*). Figure 2 shows the distributions of execution time (in hours) for this experiment.

Results. Table 3 shows detailed data for the 10 runs of AF over each configuration from A to H. The value "-" for column "CID" indicates a missed crash (resp., value "40.0" for column "time" indicates a timeout). For example, for configuration A, AF misses the crash on experiment 4. We list below our key observations:

- *Time.* The dispersion of the data points in AF is high for almost all configurations. The mean standard deviation of execution times for all configurations is 7.79h.

	Config. A		Config. B		Config. C		Config. D		Config. E		Config. F		Config. G		Config. H	
#	CID	time	CID	time	CID	time	CID	time	CID	time	CID	time	CID	time	CID	time
1	6	6.5	4	33.6	4	36.5	7	5.1	1	3.6	1	4.0	6	11.2	5	2.7
2	6	14.6	-	40.0	-	40.0	7	5.5	5	3.6	-	40.0	6	15.1	5	3.4
3	6	13.9	6	37.8	-	40.0	7	5.3	6	1.9	-	40.0	4	18.3	5	3.3
4	-	40.0	-	40.0	-	40.0	7	5.7	5	3.9	-	40.0	4	16.6	5	3.1
5	4	33.9	6	4.5	12	4.0	7	5.9	6	1.2	-	40.0	-	40.0	5	2.9
6	9	19.0	4	27.2	-	40.0	7	5.1	5	3.7	12	37.8	4	34.7	9	2.9
7	4	30.5	-	40.0	-	40.0	7	7.4	5	1.0	-	40.0	4	17.6	5	2.5
8	6	16.7	6	12.7	-	40.0	7	5.3	9	1.5	-	40.0	4	1.5	5	3.0
9	4	32.4	6	35.3	-	40.0	7	4.9	9	1.1	-	40.0	9	15.2	5	3.2
10	6	13.4	6	2.9	-	40.0	7	4.3	5	3.9	-	40.0	6	1.6	5	1.1
avg.	90%	22.1	70%	27.4	20%	36.1	**100%**	5.5	**100%**	2.5	20%	36.2	90%	17.2	**100%**	2.8

Table 3. Impact of using random seeds in AF.

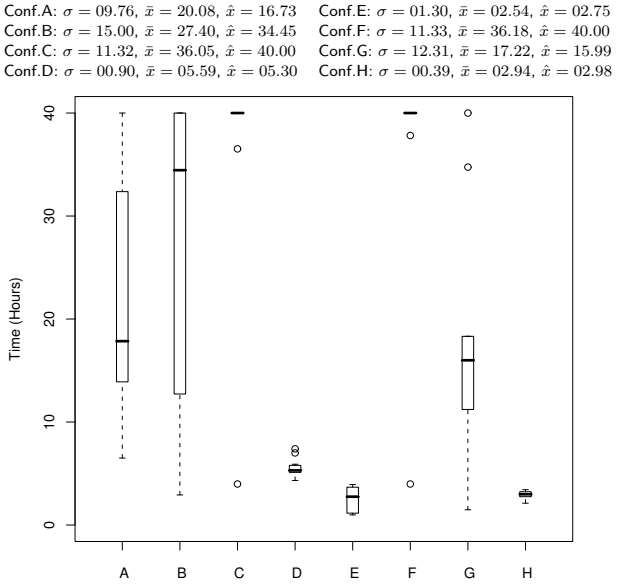

Figure 2. AF time distributions for random data.

That means that AF execution time is very sensitive to the selection of the seed.

- *Kind of crash.* Only configuration D finds the same kind of crash in all executions. All others configurations find more than 2 types of crash (for instance, Config. E finds 4 different types).

- *Precision.* Although some executions do not find a crash, AF can find crashes consistently. The mean of the precison was 74%.

3.4.2 Random sequence

This section describes the experiment we conducted to evaluate the impact of randomizing the list of *atoms* used as input to AF.

Algorithm 6: genList with Allpairs

```
1  genList(Set⟨Test⟩ suite, int nAtoms, long seed): List⟨Test⟩
2  begin genList
3      Map⟨Category, Set⟨Atom⟩⟩ partition = partition(suite);
4      Map⟨Category, Set⟨Atom⟩⟩ selected = ∅;
5      foreach entry in partition do
6          Set⟨Atom⟩ atoms = entry.value();
7          Set⟨Atom⟩ tmp = ∅;
8          for i = 1..nAtoms do tmp = tmp ∪ atoms.pickOne(seed);
9          selected.put(entry.key(), tmp);
10     endfch
       /*concatenates all sequences of atoms.  each
       sequence includes one atom on each category*/
11     return allpairs(selected);
12 end
```

Setup. We run AF for 10 times on phone configurations E and H. These configurations have the lowest average execution times (see Table 3). We used the same random seed for data and sequence generation in all runs.

Algorithm 6 redefines function *genList* that AF uses. This version associates each atom to one domain category. The categories we define are as follow: applaunch (*atoms* that only go to an application), browser (access the Internet), mms (deal with multimedia messages), multimedia (deal with multimedia files and camera), phonebook (deal with calendar, events and contacts), and sms (deal with text messages). Function *partition* in line 3 takes as input a user-defined test suite and returns a map that associates a set including all atoms of a category with the category it belongs.

The code fragment in the line range 5-10 selects *nAtoms* on each category and assigns the resulting map to variable *selected*. We apply pairwise coverage [12] to generate sequences of atoms with the property that each atom of each category is paired to another atom of another category in at least one case. For this, we use the Allpairs [1] tool. Finally, it gives as output sequences of atoms (each sequence includes one atom of each category), and then

	Config. E		Config. H	
#	CID	time	CID	time
1	-	5.5	5	1.6
2	5	2.4	-	5.0
3	5	5.3	-	5.4
4	-	10.4	5	1.3
5	-	6.2	-	6.7
6	-	4.9	5	1.1
7	5	3.1	-	4.3
8	5	6.7	-	5.5
9	-	5.4	-	5.4
10	-	5.5	5	2.2
avg.	40%	5.5	40%	3.8

Table 4. Runs of AF with different execution lists for configurations E and H

concatenates these sequences to build one longer sequence AF executes.

Results. Table 4 shows detailed data for the 10 runs of AF over configurations E and H. Our key observations for this experiment are as follows:

- *Time.* The average time Config. E (resp., Config H) took to find a crash – 5.5h (resp., 3.8h) was higher than the one for Experiment II – 2.5h (resp., 2.8h), see Table 3. However, if we only consider the runs that found a crash in Config. H, AF performed faster in this experiment. Its slowest time was 2.2h, while Experiment II reported 3.4h.
- *Precision.* For both configurations, AF found a crash in only 4 out of 10 runs, while AF found a crash for 100% of the cases when using atoms derived from SH tests.

This experiment indicates that the interaction between categories (functionalities of the system) may not be as important as the selection of critical atoms.

3.5. Impact of Randomization on BxT

This section discusses two experiments to evaluate the effect of randomization on BxT. The first experiment evaluates D-BxT compared to BxT. The second experiment evaluates the effect of using different random seeds as input to BxT.

3.5.1 Random data and sequence in BxT

This section shows the impact that the use of different random seeds has in the effectiveness of BxT.

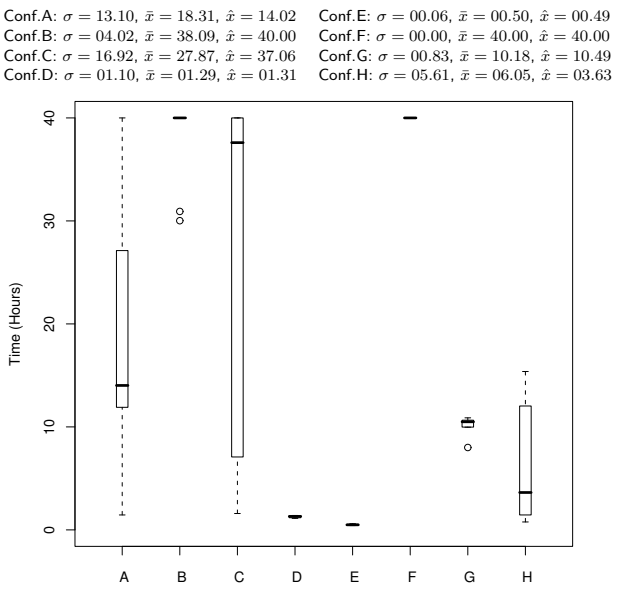

Figure 3. BXT time distributions for random data and sequence.

Setup. We ran BxT 10 times for each configuration with different random seeds. The use of different seeds impacts the generation of different sequences of events and data. With this experiment we want to observe the variance of the technique for distinct seed selections. Figure 3 shows the distribution time (in hours) and Table 5 shows the detailed data for all configurations.

Results. We list next key observations:

- *Precision.* Although BxT misses the crash in some executions, it can find crashes consistently. The mean of the precision was high (69%).
- *Variance.* The standard deviation in BxT is high for configurations A, C and H but low for the other configuration. It is likely that, for those cases, the fault density is low relative to other configurations and the selection plays an important role.

3.5.2 Comparison of BxT and D-BxT

This section compares BxT and D-BxT.

Setup. This experiment configures BxT with a timeout of 40h, and parameter *numRept* set to 1000. BxT runs 1000 events in each iteration, taking approximately 10min. D-BxT explores one screen for some time (less than 10 minutes) and jumps to another screen until it reaches the 40h timeout. The experiment configures D-BxT with the

	Config. A		Config. B		Config. C		Config. D		Config. E		Config. F		Config. G		Config. H	
#	CID	time	CID	time	CID	time	CID	time	CID	time	CID	time	CID	time	CID	time
1	6	14.6	-	40.0	-	40.0	7	1.2	5	0.5	-	40.0	6	8.0	4	12.0
2	1	1.4	-	40.0	-	40.0	7	1.4	5	0.5	-	40.0	6	10.0	4	15.4
3	-	40.0	4	39.9	1	7.1	7	1.3	5	0.5	-	40.0	6	10.6	5	1.2
4	11	12.9	-	40.0	12	32.4	7	1.3	5	0.4	-	40.0	6	10.5	5	4.5
5	6	7.8	-	40.0	1	35.2	7	1.3	5	0.5	-	40.0	6	10.7	10	0.8
6	11	11.9	-	40.0	-	40.0	7	1.1	5	0.6	-	40.0	6	9.9	4	5.8
7	11	14.0	-	40.0	12	2.5	7	1.4	5	0.5	-	40.0	6	10.5	6	2.7
8	-	40.0	4	30.0	-	40.0	7	1.4	5	0.6	-	40.0	4	10.9	5	2.7
9	1	27.7	-	40.0	-	40.0	7	1.2	5	0.6	-	40.0	6	10.7	10	1.5
10	11	14.1	-	40.0	12	1.6	7	1.4	5	0.6	-	40.0	6	10.1	5	14.0
avg.	80%	18.4	20%	38.1	50%	27.9	**100%**	1.3	**100%**	0.5	0%	40.0	**100%**	10.2	**100%**	6.1

Table 5. Impact of using random seeds in BxT.

parameter *numRept* set to 50, which results in each iteration taking approximately 30s.

Results. Table 1 summarizes the comparison. We observe the execution time difference from BxT to D-BxT considering all configuration runs, and the time difference considering only those runs that both BxT and D-BxT crash the application. We list below our key observations:

- *Time.* On average, D-BxT is slower than BxT when they both find a crash. D-BxT ran 22.6 hours more than BxT.

- *Precision.* In contrast to D-BxT and AF, BxT could not crash the application in 3 out of 8 configurations. That indicates that the screen jump was effective to improve the exploration. Conceptually, the jumps correspond to a higher weight to the width compared to the depth of the exploration graph.

3.6. Dispersion of screens in D-BxT

We conducted one experiment to evaluate whether more uniform exploration with D-BxT (i.e., exploration that visits screens with similar frequencies) correlates with time for finding a bug. The insight is that one does not want to explore for too-long regions without bugs.

We run each of the 8 phone configurations for 5 times with different seeds and measure two variables of interest: (i) dispersion of screens, and (ii) time for a bug. For dispersion, we count how many times each screen is visited (in a single exploration) and calculate the standard deviation of these counters. Figure 4 shows the scatter plot with points relating these two variables of interest. The linear regression line shows tendency. The correlation coefficient for this data set is 0.6 (ranges from 0 to 1) with a p-value (probability of rejecting the null-hypothesis) of 0.00007614. As expected the correlation between these two variables is relatively high.

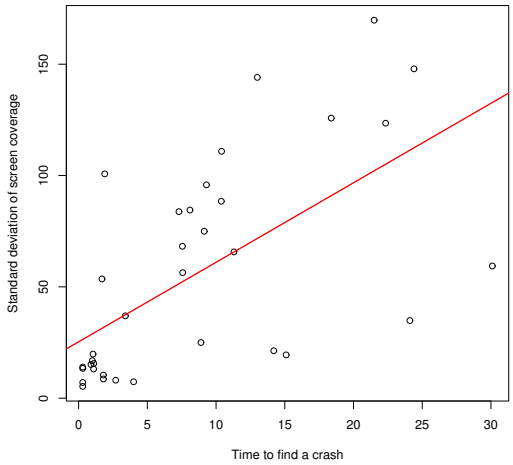

Figure 4. Correlation between dispertion of screeen counters and capability to find errors.

3.7. Threats to validity

This section describes threats to internal and external validity of our experiments. *Internal validity* determines whether the techniques have a cause-and-effect relationship in the experimental observations. *External validity* determines whether or not one can generalize the experimental observations to other scenarios.

One threat to internal validity is internal randomness. In principle, it is possible that the system does not answer promptly to the commands that the automated test issues. This depends on the operating system's scheduling decisions. This effect could impact our observations. One threat to external validity is portability of techniques. We implemented all techniques with the goal of testing cellular phones. In principle, there is no reason to believe that they

are not applicable to other kinds of application.

4. Discussion

Experimental results show that no technique subsumes the other with respect to both capability of crash and time for a crash. The techniques are therefore complementary with respect to these metrics. For instance, despite the fact that all techniques can find a crash when SH finds, SH can crash configuration H the fastest. In addition, it is very important to note that having different crash reports is very important as sometimes these crashes correspond to different faults.

5 Related and Future Work

Brooks and Memon [6] propose a technique to generate test cases based on usage information, in the form of usage profiles. These profiles describe event sequences captured from the user's experience, i.e., event sequences captured while the user interacts with the GUI. Although the idea of using profiles is appealing, it is not yet practical in the domain we evaluate our techniques (cellular communication). This work is complementary to ours.

Yuan and Memon [15] propose a technique for test case generation of GUI applications based on the analysis of feedback obtained observing the state of GUI widgets from sample executions. The goal is to include only related events in a test and therefore reduce the search space. More precisely, relate events that read (resp. write) to a part of the state that the other writes (resp. reads). Identifying these data dependencies helps significantly to reduce the search space that a test driver needs to explore. We plan to build on these ideas as future work.

Memon at al. [11] design many different generic oracles for GUI that one can use to assert her expectations of a test output. Their paper shows the importance of oracle definitions to estimate the effectiveness of the testing process. We are currently using one specific kind of oracle for detecting crashes and plan to use other kinds of oracles in the near future. More specifically, oracles to detect memory leakage and battery consumption.

6. Conclusions

This paper describes black-box testing techniques with the goal of crashing GUI. We evaluate these techniques on Motorola cellular phones with real (historical) errors.

Our empirical results demonstrate that AF and BxT together outperformed SH and DH with respect to time and also to the number of crashes reported. We also observed that the precision for AF and BxT was 74% and 69% respectively. This result indicates that a more automated technique (BxT) performed nearly the same w.r.t. precision as one using user-provided test suites as input (AF). Experimental results show that no technique subsumes the other with respect to both time and number of crashes found. In addition, different crash reports have shown to be important for identifying different bugs. This suggests that a testing team should run all algorithms when possible.

Acknowledgments. This work is in collaboration with the Stress Lab team at Motorola Brazil Test Center and is partially supported by the CNPq grants 142905/2006-2 and 550466/2005-3.

References

[1] James Bach - Satisfice, Inc webpage. `http://www.satisfice.com/tools/pairs.zip`.

[2] Linux Java webpage. `http://www.motorola.com/motomagx/`.

[3] Symbian webpage. `http://www.symbian.org`.

[4] L. Apfelbaum and J. Doyle. Model based testing. In *Software Quality Week Conference*, pages 296–300, 1997.

[5] B. Beizer. *Software Testing Techniques*. International Thomson Computer Press, 1990.

[6] P. A. Brooks and A. M. Memon. Automated gui testing guided by usage profiles. In *ASE '07*, pages 333–342, New York, NY, USA, 2007. ACM.

[7] A. Gotlieb. Exploiting symmetries to test programs. In *ISSRE*, Denver, Colorado, November 2003.

[8] M. Harman and J. Wegener. Getting results from search-based approaches to software engineering. In *ICSE*, pages 728–729, 2004.

[9] R. Iosif. Exploiting heap symmetries in explicit-state model checking of software. In *ASE*, page 254, Washington, DC, USA, 2001. IEEE Computer Society.

[10] D. Lee and M. Yannakakis. Principles and methods of testing finite state machines - A survey. In *Proceeding of The IEEE*, volume 84, pages 1090–1123, Aug. 1996.

[11] A. M. Memon, I. Banerjee, and A. Nagarajan. What test oracle should I use for effective GUI testing? In *ASE*, pages 164–173, 2003.

[12] G. J. Myers. *Art of Software Testing*. John Wiley & Sons, Inc., 1979.

[13] P. Santhanam and B. Hailpern. Software debugging, testing, and verification. *IBM Systems Journal*, 41:4–12, 2002.

[14] W. Visser, C. S. Pasareanu, and R. Pelanek. Test input generation for Java containers using state matching. In *ISSTA*, pages 37–48, 2006.

[15] X. Yuan and A. M. Memon. Using gui run-time state as feedback to generate test cases. In *ICSE*, pages 396–405, Washington, DC, USA, 2007. IEEE Computer Society.