

Feature Matching-based Approaches to Improve the Robustness of Android Visual GUI Testing

Original

Feature Matching-based Approaches to Improve the Robustness of Android Visual GUI Testing / Ardito, L., Bottino, A., Coppola, R., Lamberti, F., Manigrasso, F., Morra, L., Torchiano, M.. - In: ACM TRANSACTIONS ON SOFTWARE ENGINEERING AND METHODOLOGY. - ISSN 1049-331X. - 31:2(2022), pp. 1-32. [10.1145/3477427]

Availability:

This version is available at: 11583/2915055 since: 2021-11-24T17:03:41Z

Publisher:

ACM

Published

DOI:10.1145/3477427

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

ACM postprint/Author's Accepted Manuscript

© ACM 2022. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in ACM TRANSACTIONS ON SOFTWARE ENGINEERING AND METHODOLOGY, <http://dx.doi.org/10.1145/3477427>.

(Article begins on next page)

Feature Matching-based Approaches to Improve the Robustness of Android Visual GUI Testing

LUCA ARDITO, Politecnico di Torino
ANDREA BOTTINO, Politecnico di Torino
RICCARDO COPPOLA, Politecnico di Torino
FABRIZIO LAMBERTI, Politecnico di Torino
FRANCESCO MANIGRASSO, Politecnico di Torino
LIA MORRA, Politecnico di Torino
MARCO TORCHIANO, Politecnico di Torino

In automated Visual GUI Testing (VGT) for Android devices, the available tools often suffer from low robustness to mobile fragmentation, leading to incorrect results when running the same tests on different devices.

To soften these issues, we evaluate two feature matching-based approaches for widget detection in VGT scripts, which use, respectively, the complete full-screen snapshot of the application (*Fullscreen*) and the cropped images of its widgets (*Cropped*) as *visual locators* to match on emulated devices.

Our analysis includes validating the portability of different feature-based visual locators over various apps and devices and evaluating their robustness in terms of cross-device portability and correctly executed interactions. We assessed our results through a comparison with two state-of-the-art tools, EyeAutomate and Sikuli.

Despite a limited increase in the computational burden, our Fullscreen approach outperformed state-of-the-art tools in terms of correctly identified locators across a wide range of devices and led to a 30% increase in passing tests.

Our work shows that VGT tools' dependability can be improved by bridging the testing and computer vision communities. This connection enables the design of algorithms targeted to domain-specific needs and thus inherently more usable and robust.

CCS Concepts: • **Software and its engineering** → **Software defect analysis; Software verification and validation; Software testing and debugging**; • **Computing methodologies** → *Matching*.

Additional Key Words and Phrases: Mobile Computing, Software Testing, Visual GUI Testing, Feature matching

ACM Reference Format:

Luca Ardito, Andrea Bottino, Riccardo Coppola, Fabrizio Lamberti, Francesco Manigrasso, Lia Morra, and Marco Torchiano. 2018. Feature Matching-based Approaches to Improve the Robustness of Android Visual GUI Testing. *ACM Trans. Softw. Eng. Methodol.* 1, 1 (July 2018), 33 pages. <https://doi.org/10.1145/1122445.1122456>

Authors' addresses: Luca Ardito, Politecnico di Torino, Corso Duca degli Abruzzi 29, Torino, Italy, 10129, luca.ardito@polito.it; Andrea Bottino, Politecnico di Torino, Corso Duca degli Abruzzi 29, Torino, Italy, 10129, andrea.bottino@polito.it; Riccardo Coppola, Politecnico di Torino, Corso Duca degli Abruzzi 29, Torino, Italy, 10129, riccardo.coppola@polito.it; Fabrizio Lamberti, Politecnico di Torino, Corso Duca degli Abruzzi 29, Torino, Italy, 10129, fabrizio.lamberti@polito.it; Francesco Manigrasso, Politecnico di Torino, Corso Duca degli Abruzzi 29, Torino, Italy, 10129, francesco.manigrasso@polito.it; Lia Morra, Politecnico di Torino, Corso Duca degli Abruzzi 29, Torino, Italy, 10129, lia.morra@polito.it; Marco Torchiano, Politecnico di Torino, Corso Duca degli Abruzzi 29, Torino, Italy, 10129, marco.torchiano@polito.it.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2018 Association for Computing Machinery.

1049-331X/2018/7-ART \$15.00

<https://doi.org/10.1145/1122445.1122456>

1 INTRODUCTION

Modern Android apps have reached a high complexity level, almost bridging the gap with traditional desktop software in terms of exhibited features. Nowadays, mobile apps are using sophisticated frameworks and modern development processes, with quick delivery cycles. Moreover, there is a very tight competition on the online markets where they are released: the official Google Play store counts 2.96 million apps available as of June 2020, and 84.3 billion downloads for the whole 2019 [53]. These characteristics should encourage a thorough Verification and Validation phase to gain the necessary confidence that the apps behave correctly and systematically use automated techniques to support and complement the slow and error-prone manual practice.

The vast majority of Android apps are Graphical User Interface (GUI) intensive and collect most of the user's interaction through GUI widgets, or *Views*. Thereby, a critical issue is ensuring that the running app correctly renders the widgets. Recent years witnessed the development of many End-to-End (E2E) testing tools that allow developers to create repeatable test scripts mimicking a human interaction with the finished app and evaluate their response [12]. Many of these tools identify the GUI widgets through layout properties of the GUI structure that serve as textual *locators* and are hence called *Layout-based* testing tools. Similarly, the execution correctness is validated by verifying the properties of the GUI layout. These tools, however, are unable to test the actual graphical appearance of the Application Under Test (AUT) when shown to its final user. Thus, visual failures may slip through undetected.

To address this limitation, researchers proposed to tackle app testing with the Visual GUI Testing (VGT) paradigm [2]. With this paradigm, the verification of behavior's correctness involves a visual comparison between the app's current and expected visual appearance. This comparison leverages image recognition algorithms and uses screen snapshots as both *visual locators* (to identify the widgets) and *oracles* (to determine whether the displayed widget is correct). Thus, the main advantage of the VGT approach is that it allows testers to verify that the widgets are displayed in the correct position and appearance. In contrast, layout-based techniques have a limited ability to verify the rendered GUI's appearance and typically only check a widget's instantiation and not whether its appearance is correct.

In addition to the visual verification component, VGT techniques allow emulating user interaction with the GUI in an entirely agnostic way to AUT implementation, operating system (OS), and platform. These features make VGT techniques optimal for testing those applications that need to provide high portability across different platforms. In particular, they represent a valid alternative to the (costly) manual testing of those applications that contain a lot of visual content, whose rendering must be carefully verified. Although the demonstrated benefits of applying VGT tools for desktop software [3], in the industrial practice of mobile app development, the inherent characteristics of the domain have slowed down their adoption. Indeed, visual tests of Android apps are very fragile due to hardware fragmentation [31]. Therefore, since every app must be compatible with many different devices (with varying screen sizes, pixel densities, and rendering specifications), marginal variations in the graphical rendering can invalidate the recognition of visual locators and oracles. As a result, test cases may not be portable to devices different than those on which the captures have been performed. Moreover, graphical changes in the same app's consecutive releases may break the tests, requiring additional maintenance in existing test suites to adapt locators and oracles. Our previous studies have shown that state-of-the-art VGT testing tools are complicated to port to different devices unless leveraging hybrid techniques that regenerate VGT tests for each device from an original layout-based test suite [22]. However, such an approach increases the test suite maintenance costs since visual locators' regeneration is a time-consuming operation that must be performed at each new application release that includes changes in the GUI appearance.

99 With this discussion in mind, one possible question is how to improve the effectiveness of the
100 VGT paradigm. The hypothesis underlying our work is that the algorithms currently used by
101 state-of-the-art VGT tools (which usually leverage pixel-to-pixel comparisons) can be made more
102 robust to possible graphical variations across devices, OSs, and versions. In particular, our work
103 aims at analyzing the role of feature matching algorithms, a specific class of Computer Vision (CV)
104 algorithms, as an enabling technology for more robust Android VGT tools [59]. These algorithms
105 rely on the analysis of local textural features invariant to several image variations (like color, shape,
106 scale, and rotation), allowing them to detect one or more targets in cluttered scenes robustly.

107 The idea of applying feature matching algorithms in the VGT domain is not new [1, 48, 58].
108 However, to the best of our knowledge, the approach presented in this paper represents the first
109 attempt to systematically evaluate the suitability of different feature matching algorithms to support
110 VGT applications and, specifically, their robustness to device fragmentation. In particular, we present
111 the design of two different feature matching-based algorithms to identify widgets' visual locators,
112 referred to in the following as *Fullscreen* and *Cropped*. Our experimental results show how our
113 approach can increase the robustness of visual locator matching, improve over state-of-the-art VGT
114 tools, and achieve higher portability across devices.

115 To validate the robustness of visual locator matching, we perform a two-fold analysis. First,
116 we validate the portability of different locator matching strategies and feature descriptors on
117 different devices over a large set of Android apps by comparing recall, precision, and execution
118 time. Second, we perform an exploratory study on a real test suite, and we compare our feature
119 matching approaches with state-of-the-art VGT tools in terms of resilience to device change.

120 Additionally, we contribute to state of the art in the field by introducing DatAndroid¹, a dataset
121 including screenshots and associated metadata (i.e., View Hierarchies) of about 100 apps emulated
122 on 14 different devices, for a total of roughly 1,400 combinations, which could serve as a public
123 benchmark for the portability of VGT techniques. To the best of our knowledge, the only available
124 public large-scale datasets with screenshots are RICO and ERICA [26]. However, they do not
125 provide screenshots of the same app rendered on different devices, with varying screen sizes,
126 pixel densities, and rendering specifications; hence, they are not suitable for investigating the
127 performance, robustness and portability of VGT tools, or their underlying CV techniques.

128 The present manuscript is structured as follows: Section II provides background information
129 about Android GUI Testing techniques, VGT tools, and CV techniques for GUI testing; Section III
130 introduces the two matching algorithms that we propose; Section IV illustrates the experimental
131 procedure to evaluate feature matching algorithms for VGT; Section V provides a description and
132 discussion of the collected results; Section VI summarizes the threats to the validity of the current
133 study and, finally, Section VII concludes the paper and provides directions for future work.

134 2 BACKGROUND AND RELATED WORK

135 In this section, we first illustrate the main concepts behind automated GUI testing, then we examine
136 the Android-specific VGT techniques and we highlight their major limitations. We finally provide
137 an overview of recent literature related to image recognition algorithms' application to VGT for
138 Android apps.
139

140 2.1 Android GUI structure and automated GUI testing

141 Android apps (and mobile apps, in general) can be divided into three main categories, according
142 to the way they are built: *native* apps, created using the Android-specific design guidelines and
143 the related development framework; *web-based* apps, i.e., web applications optimized to be loaded
144

145
146 ¹The dataset is available for download on Zenodo or at the link <https://frankissimo.github.io/datAndroid/>
147

148 on browsers installed on Android devices; *hybrid* apps, combining the principles of native and
149 web-based apps by employing components that are capable of loading dynamic content from the
150 web at run time.

151 While web-based apps and the content of hybrid apps are designed by following the usual web
152 development principles, native apps are built with a precise set of visual components. Each app
153 screen is managed by an *Activity*. This component defines and builds the GUI shown to the user,
154 which is composed of Views (i.e., widgets) arranged in a tree structure according to a specific layout
155 [52]. The layout can be defined programmatically (in the code of the Activity class) or by defining
156 static XML files that can be loaded (or “inflated”) as the Activity’s first operation. Screen transitions
157 are obtained by registering callbacks on interactable GUI elements (like buttons and text fields) and
158 processing user inputs (like clicks, text insertions, and swipe operations).

159 One of the most widely used approaches to automated GUI testing for Android applications is
160 layout-based testing. This approach leverages the possibility to extract, at any given time during the
161 app execution, the current screen hierarchy, where each View is associated with a set of platform-
162 specific parameters, like unique id, textual content, size, and screen position. These properties allow
163 identifying the visible elements used by Layout-based testing tools as both locators (to identify the
164 Views to interact with) and oracles (to verify if a specific View has been shown on the screen with
165 the desired properties).

166 The relevance of Layout-based testing is witnessed by the fact that most of the research in
167 automated GUI testing for Android apps focused on this approach. Linares-Vasquez et al. [38]
168 provided a classification of over 80 testing tools based on the approach adopted to define the
169 test sequences. According to this classification, *Automation Frameworks and APIs* allow testers to
170 create JUnit-like test scripts to be run against Android GUIs. Examples of this category are the
171 tools embedded with the Android development framework (Espresso [50], UIAutomator [64]), and
172 Robotium [63]). *Record and Replay* tools create sequences of interactions by capturing manual
173 sequences executed by a tester into repeatable scripts. Recent examples are Barista [28] and
174 OBDR [46]. *Automated Test Input Generation Techniques* execute tests without the need for manually
175 defining input sequences; the sequences can be generated randomly, as with SAPIENZ [42], or
176 Monkey [27], the official GUI random exerciser provided by Android. More sophisticated approaches
177 automatically create the test sequences by generating and using GUI models, like the finite state
178 machines exploited by MobiGUITAR [7] and DroidMate [15].

179 2.2 Tools for Visual GUI Testing

181 VGT is an alternative approach to GUI testing that uses CV to automate testers’ tasks [19]. Test cases
182 developed by the VGT approach show several key features: high readability (since each interaction
183 is described with screen captures of the application and not with code); a more natural development
184 of scripts with respect to the Layout-based approach (which, for the definition of each locator,
185 requires the tester/developer search unambiguous properties in the layouts); complete platform
186 independence (since the GUI tests are agnostic of the specific platform and OS for which the AUT
187 is engineered, and they only require to render or emulate it). Empirical controlled experiments
188 have demonstrated that VGT tools can significantly increase testers’ productivity [10].

189 Several commercial and open-source tools have adopted the VGT approach. Sikuli [62] uses image
190 recognition algorithms to identify GUI components. The tool provides an integrated development
191 environment (IDE) for assisted script generation and can output test scripts in Python, Ruby, and
192 JavaScript. Libraries for using Sikuli matching commands inside JUnit test cases are also available.
193 EyeAutomate (evolution of JAutomate [6]) uses a proprietary algorithm to recognize graphical
194 oracles and provides an IDE for manual script definition as well as a Java library to write down
195 JUnit-like test methods. The tool supports the creation of text-based repeatable scripts written in a
196

197 proprietary syntax. Layout Bug Hunter (LBH) validates bugs in the GUI rendering by checking
198 layout errors caused by changing the screen size [32]. Commercial products are also available, like
199 AppliTools, which uses visual inputs combined with machine learning capabilities, and PhantomCSS,
200 based on pixel-by-pixel screen comparisons [60].

201 Several studies in the literature proved the applicability of these tools in real-world scenarios.
202 Alegroth et al. documented the feasibility of the VGT technique's long-term usage through an
203 industrial case study [3]. Borgesson et al. performed comparative studies between Sikuli and
204 JAutomate in an industrial setting, finding that VGT is an applicable technology for automated
205 system testing with significant advantages compared to manual system testing and manual scripting
206 and proving that the technique can be successfully used for automated acceptance testing [16].

207 However, as highlighted by most VGT studies, the main limitation of this approach is the high
208 maintenance cost of visual test suites [4]. Several studies tried to address the maintenance issues by
209 providing translation-based or hybrid approaches that combine visual and Layout-based locators'
210 characteristics by using the latter to automatically re-generate the visual captures [5, 11, 36].
211 These approaches are however limited to the translation of a limited set of types of locators and
212 interactions, and generally do not achieve a 100% precision in visual oracle generation, thereby
213 requiring manual intervention at times. They also require the layout-based test suite to be rendered
214 against each emulated device for the creation of visual screen captures, therefore they are not
215 indicated for large-sized test suites.

216

217 2.3 Issues with VGT of Android apps

218 Beyond the limitations introduced in the previous section, Android fragmentation is recognized as
219 a relevant issue to tackle by developers and testers [42]. The fragmentation concept can be divided
220 into *hardware fragmentation*, i.e., inconsistencies among various hardware specifications that can
221 cause differences in GUI renderings, and *software fragmentation*, i.e., inconsistencies due to the
222 different versions of the OS, application programming interface (API), and GUI customizations
223 [51]. Fragmentation mandates additional effort by both developers and testers to ensure that their
224 apps' features are entirely portable to the devices that the app must provide compatibility with. A
225 comprehensive testing procedure shall test the app on multiple handsets, each with its hardware
226 and software characteristics [34].

227 Fragmentation significantly impacts GUI visual test practices since it results in the impossibility
228 to guarantee that the image recognition algorithms can correctly find the locators and oracles
229 captured on a source device. A further issue is that the Android framework allows rendering
230 layouts in different ways (i.e., with a variable number and arrangement of widgets) on devices with
231 different pixel densities and screen sizes. For instance, multiple widgets can be compacted in menus
232 or inserted into scroll views on smaller screen sizes and visualized only after swipe operations
233 (as in figure 1, which reports an example of the same Activity rendered on two different devices).
234 Thereby, GUI test cases are intrinsically fragile to fragmentation, since even if no faults are present
235 when executing the app on different devices, visual test scripts may require the locators to be
236 adapted to the specific size and arrangement of the widgets on the target device to avoid wrong
237 test executions.

238 Another major problem documented for VGT, although not exclusive to the Android domain,
239 is the *graphical fragility*, driven by pure graphical changes to widgets or modifications in AUT
240 layout properties (e.g., textual content or hierarchy changes) [24]. These situations require test
241 maintenance, forcing testers to provide new visual locators for each widget whose appearance has
242 changed. Controlled experiments on mobile applications have measured that maintenance effort
243 due to graphical fragilities in VGT test suites can account for up to 30% of total test maintenance
244 time [21].

245

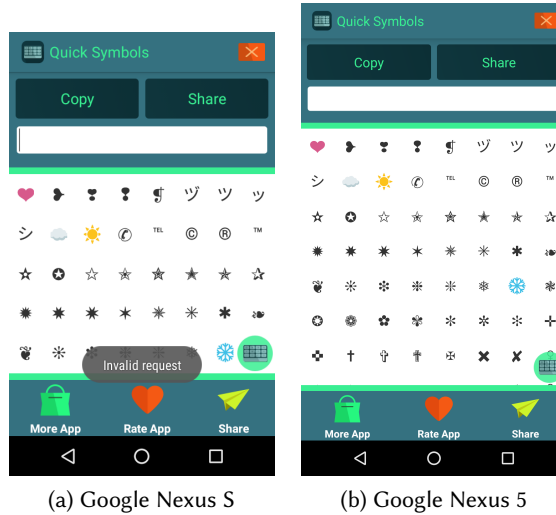


Fig. 1. Main Activity of the ASCII Notes application on devices with different screen sizes. On the smaller Nexus S screen, a lower number of components is displayed than on the Nexus 5 screen.

A final drawback of VGT tools is that they are commonly run in a desktop environment that emulates the AUT on an Android Virtual Device (AVD). This approach adds an additional layer of fragility to visual tests since the recognition of the locators is influenced by the rescaling operated on the AVD by the host OS [11][37].

2.4 Computer Vision techniques for automated GUI testing

Several VGT tools like Sikuli [2] and EyeAutomate [1] rely on CV techniques to automate the search of visual locators. The most common approach, which is also our work's goal, is to recognize a visual locator's location in the device screen capture using image matching algorithms.

The literature has documented several failures and issues associated with commercial and open-source VGT tools. One of the main difficulties in analyzing the causes of these problems is that the algorithms' full specifications are often undisclosed. Nonetheless, some data on the matching algorithms' performance (mainly exploiting their implementation in the OpenCV library) is available in the literature [17, 47, 48].

Among the available VGT tools, Sikuli is based on a simple *template matching* technique that searches the visual locator position in the target screenshot. The same approach is also discussed in [47]. This algorithm simply slides the template image over the target screenshot and compares the template and the underlying target patch. This comparison involves a similarity metric such as the squared difference or the normalized correlation of pixel intensities. The sliding approach of template matching potentially requires sampling a large number of points. Thus, VGT tools typically reduce the search space's size by stopping as soon as a suitable match is found. However, this choice may be suboptimal if the same or similar widgets occur within the application. Moreover, template matching is not robust to variations in scale (which occur when the image is rendered on devices with different physical characteristics) and rotation and color (which are relatively frequent during the evolution of the AUT [49]), leading to wrong executions due to image recognition failures.

295 More recently, matching algorithms based on local feature descriptors emerged as a more effective
296 and robust alternative to pixel-based similarity. Feature matching refers to a set of CV techniques
297 that aim to find corresponding or similar image patches in two images, with numerous applications
298 including stereo matching and image retrieval [25]. A feature matching algorithm comprises three
299 essential steps: *keypoint extraction*, *local feature extraction*, and *keypoint matching*.

300 The goal of keypoint extraction is to find distinctive locations that can increase the uniqueness
301 of the features extracted, usually promoting high-contrast regions of the image, such as object
302 corners. These keypoints are then associated with a *local feature vector*, i.e., a local texture *descriptor*
303 that provides a strong characterization of the local image patches surrounding an image point,
304 invariant to several image variations (like color, shape, scale, and rotation). Several local feature
305 descriptors have been defined [39, 59]. Popular choices include SIFT [39], SURF [14], and Akaze
306 [29], which differ in terms of computational cost, accuracy, and robustness [54]. Finally, keypoints
307 extracted from the two images are matched using suitable metrics in the feature space (i.e., matched
308 keypoints identify visually similar regions with a low distance in the feature space) to establish a
309 point-to-point correspondence between the two images.

310 An example of keypoint descriptors' application to VGT is MAuto [58], a tool that generates test
311 sequences for mobile games through the Record and Replay technique. MAuto leverages AKAZE
312 feature descriptors to identify locators in the AUT. However, the excessive number of features in
313 the query images used to identify locators limited the approach's applicability. As a workaround to
314 this issue, the authors restricted the search to a limited region of the query image, which had to be
315 appropriately and heuristically calibrated to provide a suitable number of features for identifying
316 the locators.

317 Available studies related to feature matching algorithms applied to the VGT tasks suffer from
318 two main limitations. The first is algorithmic in nature, as most of the state-of-the-art tools and
319 approaches take as input only the cropped image of a single input visual locator [1, 2, 58]. This
320 approach is not robust in the presence of repeated or similar widgets, where the locator could be
321 potentially matched to multiple widgets. To address this issue, we propose a Fullscreen matching
322 algorithm which leverages all visual information available in the source and target screenshots,
323 and compare it experimentally with the state-of-the-art approach. The second limitation lies in
324 the limited experimental validation, which is usually based on a few selected apps or case studies
325 [58]. Previous studies therefore do not provide a comprehensive assessment of state-of-the-art CV
326 algorithms. To close the gap, we offer to the community DatAndroid, a dataset of roughly 100 apps
327 rendered on 14 different devices, as well as an experimental methodology to systemically compare
328 the robustness of different algorithms and descriptors in the presence of device fragmentation.

329 3 FEATURE MATCHING BASED VGT

330 As mentioned in the introduction, our primary goal was to design and develop an approach capable
331 of (i) matching widgets rendered on Android app screens, (ii) performing interactions on individual
332 widgets, and (iii) running test sequences on them. The main constraint on our method is that it
333 must locate the widgets of interest on various devices where the app may run. In the following, we
334 adopt these definitions:
335

- 336 • *Source device*: the device where the visual locator is captured;
- 337 • *Target device*: the device where the locator should be found.

338
339 Our purely visual testing approach uses only rendered screens as input and, differently than
340 several state-of-the-art applications of VGT to Android apps, we search for the visual locators
341 directly in the screen captures obtained by ADB (Android Debug Bridge) interaction with the
342 Android devices rather than in the renderings of the screen on the host desktop system where the
343

Table 1. Input and output of the two algorithms

	Fullscreen algorithm	Cropped algorithm
Input / Locator	fullscreen capture of the source device screen, bounding box of the widget	Cropped screen capture containing only the widget
Result	x, y coordinates belonging to the widget in the target device screen	

device is emulated. This approach's advantages are twofold: it can be applied to both emulated and real devices, and it does not need to consider the resizing of the virtual device GUI operated by the host OS. We also emphasize the fact that, in principle, the source and target devices may coincide. This condition occurs, for example, when comparing different versions of GUIs that have graphical variations.

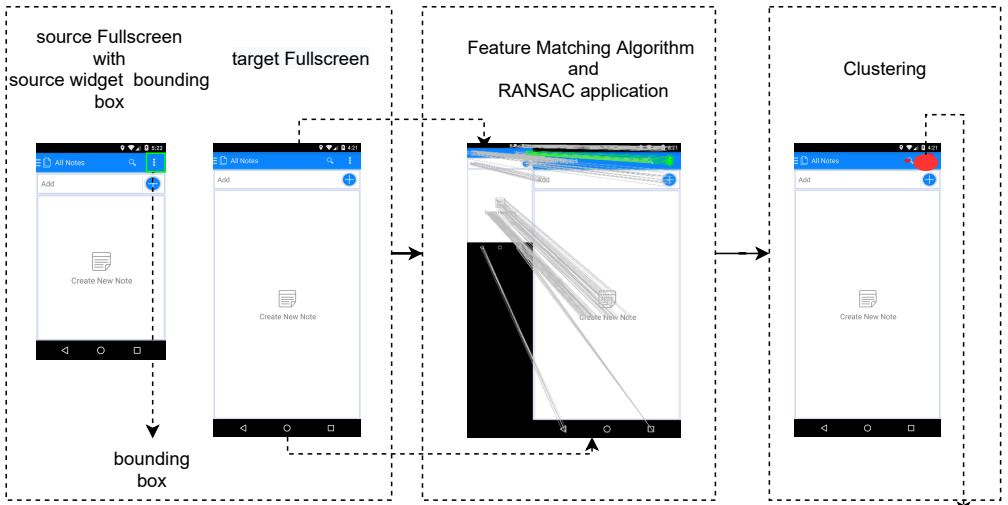
The proposed feature matching based VGT is based on determining, for each visual locator in the source screenshot, its location in the target image (or lack thereof). This analysis leverages feature matching algorithms, which identify in the target the location of a region of the source image that represents the widget to be searched. Under these conditions, the matching algorithms can be handled in two different ways, referred to in the following as *Fullscreen* and *Cropped* (summarized in Table 1). The main steps of both algorithms are illustrated and compared in Figure 2.

Assuming for simplicity that the test involves a single widget, the visual locator on the source device (e.g., the bounding box of the widget) is identified in both cases by a human tester during the creation of the test suite. The main difference between the two algorithms is that, in *Fullscreen*, we use for identification all visual information available in the source screenshot, whereas, in *Cropped*, we use only the information present in the visual locator.

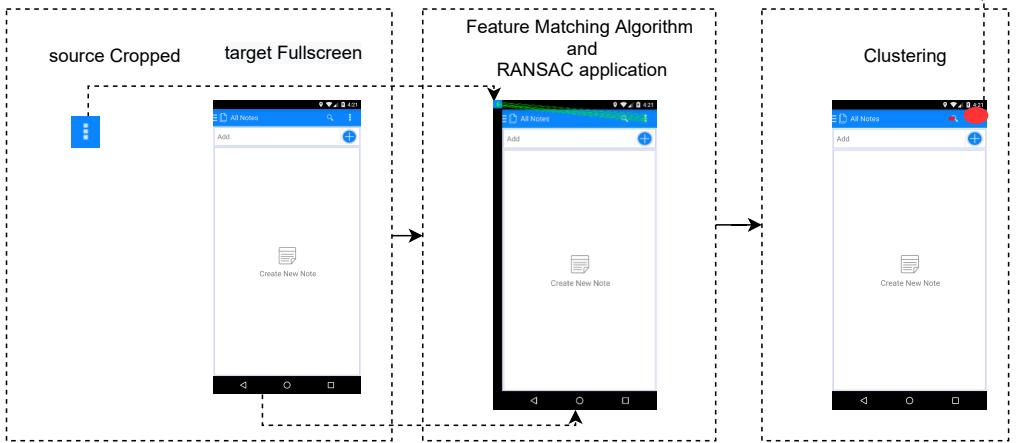
The identification of the position of the visual locator is based on established feature matching techniques. These techniques leverage local texture descriptors (see Section 2.4 for details and Section 4.3.2 for the description of the specific descriptors analyzed in our research). For tasks like image registration, object tracking, and recognition, robust matching algorithms can be applied to pair keypoint descriptors obtained from the source and target images. These algorithms work as follows. First, the best match between descriptors is computed by means of a suitable metric in the feature space (e.g., euclidean distance). Then, since the extracted feature point pairs might suffer from significant correspondence errors or mismatches in the pairing, a common strategy is to postprocess candidate matches with robust data fitting techniques such as RANSAC (which stands for RANdom SAMple Consensus) [13]. RANSAC starts from an initial estimate of the homography relating the two images (computed from all the matches) and then iteratively removes the outliers, i.e., the matches that are not consistent with the estimated homography, to update the homography with the remaining set of inliers. Figure 3 shows an example of matching a source and target snapshot, where the initial keypoint matches (Figure 3 left) are "cleaned" applying RANSAC as a post-processing step (Figure 3 right). The remaining matches will be the base of the clustering method used to identify the locator position in the target image.

In the following sections, we detail the *Fullscreen* and *Cropped* algorithms where, for the sake of clarity and without loss of generalization, we assume again that the test involves a single widget. Implementation details are reported in Section 4.3.2.

Fullscreen algorithm



Cropped algorithm



(X,Y)
interaction coordinate

Fig. 2. Schematic representation of the two matching algorithms compared in this work. The example shows the AcNotes app with Nexus S and Pixel 3a XL as source and target devices, respectively. First row: *Fullscreen*. From left to right: the source and target screenshot (the green region indicates the visual locator); results of the keypoint matching (green lines indicate the matched keypoints associated to the source widget, grey lines the other matched keypoints); the identified interaction coordinates (identified by the larger red circle). Second row: *Cropped*. From left to right: the cropped widget and target screenshot; results of the keypoint matching; the identified interaction coordinates (identified by the larger red circle). Both algorithms match the source widget with two potential target widgets, identified by the clustering step as two clusters of keypoints (highlighted by the small and large red circles); the largest one is selected as interaction coordinates.

442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490

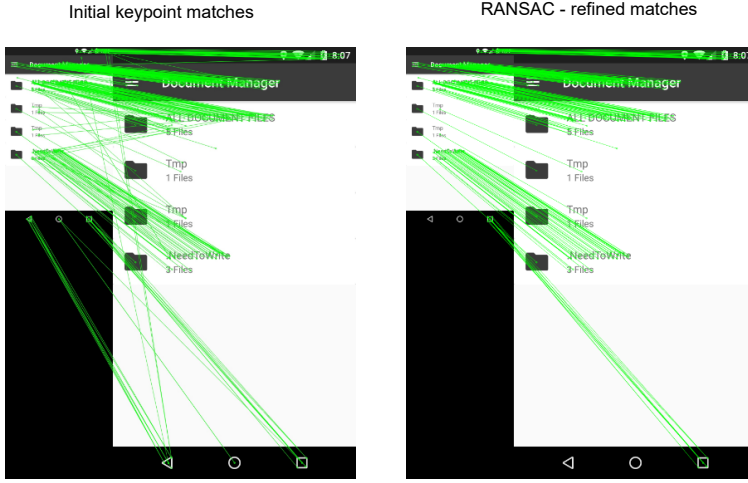


Fig. 3. Example of RANSAC post-processing (in Fullscreen algorithm). Left: Initial keypoint matches. Right: RANSAC-refined matches. In both images, the green lines indicate the matched keypoints.

3.1 Fullscreen matching algorithm

The Fullscreen matching algorithm (Algorithm 1) is summarized in the top part of Figure 2. It receives as input both the source and target fullscreen images, along with the (manually identified) bounding box of the source locator.

With this approach, we first extract the keypoints from both source and target images, and then we post-process their matches with RANSAC. To identify the widget's target location, we use the widget bounding box to prune the set of matches (i.e., by discarding all matches whose source keypoints fall outside the bounding box). Finally, to obtain the most likely interaction coordinate pairs (i.e., the widget interaction coordinates on both source and target, which are needed by the test procedure), we apply the MeanShift clustering algorithm [30] to the source and target keypoints. On each device, the largest cluster found identifies the target locator, and its centroid is returned as the final output of the matching algorithm.

Algorithm 1 Fullscreen matching algorithm

- 1: $FsSource \leftarrow$ Source image of Full Screenshot
 - 2: $FsTarget \leftarrow$ Target image of Full Screenshot
 - 3: $Src_bb_target \leftarrow$ Source bounding box
 - 4: **function** FULLSCREEN($FsSource, FsTarget, Src_bb_target$)
 - 5: $src_pts, targ_pts \leftarrow$ feature_matching($FsSource, FsTarget$)
 - 6: $src_pts, targ_pts \leftarrow$ RANSAC($src_pts, targ_pts$)
 - 7: $src_pts, targ_pts \leftarrow$ select_target_keypoints($src_pts, targ_pts, Src_bb_target$)
 - 8: $click_coordinates \leftarrow$ Meanshift($targ_pts$)
 - return** $click_coordinates$
-

The advantages of the Fullscreen algorithm are two-fold. First, it reduces the time needed to process the entire test suite if more than one widget has to be extracted for the test generation. Second, RANSAC relies on the computation of a homography that implicitly favors spatial widget

491 arrangements that are similar between source and target. Therefore, when the same (or similar)
 492 widgets are repeated in the GUI, it is possible to match each widget with its most similar counterpart,
 493 with RANSAC ensuring that the final matching is also spatially coherent. A typical example that can
 494 benefit from these characteristics is a calendar application, where the same numeric symbols repre-
 495 sent different days of the month. However, in these situations, RANSAC may occasionally remove
 496 a correct match. The main disadvantage of this algorithm, however, is the higher computational
 497 cost than Cropped when a single widget has indeed to be identified.

498 3.2 Cropped matching algorithm

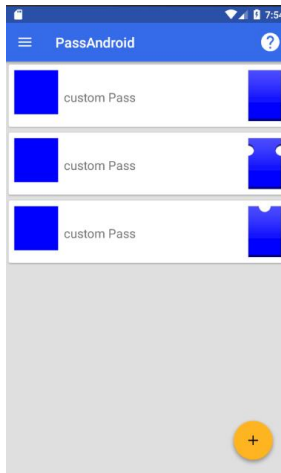
500 The Cropped matching algorithm (Algorithm 2) is summarized in the bottom part of Figure 2. It
 501 receives as input the cropped image of the source device widget and the target screenshot (i.e., the
 502 same input required by Sikuli and other comparable tools). The source keypoints are extracted
 503 only from this cropped image and matched to those extracted from the target screenshot. Again,
 504 we apply RANSAC for outlier removal and MeanShift for the identification of the interaction
 505 coordinates.

506 Algorithm 2 Cropped matching algorithm

```

507 1:  $CsSource \leftarrow$  Source image of Cropped Screenshot
508 2:  $FsTarget \leftarrow$  Target image of Full Screenshot
509 3: function CROPPED( $CsSource, FsTarget$ )
510 4:    $src\_pts, targ\_pts \leftarrow$  feature_matching( $CsSource, FsTarget$ )
511 5:    $src\_pts, targ\_pts \leftarrow$  RANSAC( $src\_pts, dst, targ\_pts$ )
512 6:    $click\_coordinates \leftarrow$  Meanshift( $targ\_pts$ )
513   return  $click\_coordinates$ 
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532

```



533 Fig. 4. Sample content for the main Activity of PassAndroid with repeated visual locators in the same screen
 534 (e.g., the “custom Pass” text box, and the blue squares on the left).

535
 536 This algorithm’s main advantage is that it is faster than Fullscreen when the test involves a single
 537 widget. However, there can be identification errors when the source locator is present multiple times
 538 in the target screen (as in the example in Figure 4, which shows a screenshot of the PassAndroid
 539

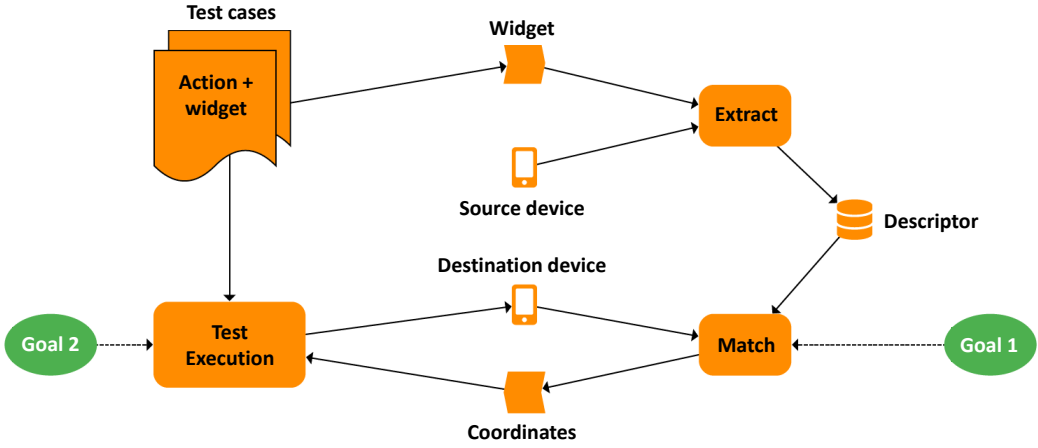


Fig. 5. Visual experimental design summary.

Table 2. Goal-Question-Metric template for the study

	Goal 1	Goal 2
Object of Study	Feature matching algorithms	Visual testing
Purpose	Compare different descriptors and matching algorithms	Compare with state-of-the-art VGT tools
Focus	Precision, recall, execution time	Device fragmentation, working tests, portability
Context	Matching of mobile GUI components	Mobile GUI test suites
Stakeholders	Researchers	Researchers, developers, testers

application that contains repeated graphical and textual content in the same Activity). In these cases, the matching algorithm returns at most one locator, which is not necessarily the correct one.

4 EXPERIMENTAL DESIGN

Our experimental assessment, summarized in Figure 5, addresses two different aspects, the feature matching algorithms and the visual testing process.

4.1 Goals

We report the Goal-Question-Metric (GQM) [18] template for the study in Table 2.

The first goal aims at assessing the effectiveness of feature matching algorithms in identifying widgets on Android GUIs. We compared different algorithms and descriptors based on standard performance measures for classification and retrieval algorithms (recall, precision) and execution time. The results are then interpreted according to researchers' perspective in the CV field, providing evidence about the performance of such techniques in the domain of Android GUIs. The aim is to enable selecting the combination of algorithm and descriptor with the highest performance on a wide range of applications, widgets, and source/target devices.

589 The second goal concerns feature matching algorithms' applicability for recognizing visual
590 locators and oracles in visual test suites for Android applications. We compared the feature matching
591 technique with state-of-the-art VGT tools by considering the impact of device fragmentation on test
592 scripts. The results pertaining to this goal are then interpreted according to researchers' perspectives
593 in the GUI testing field, testing tool creators, and developers. Although, in theory, we expect that a
594 feature matching technique with higher performance can enhance test portability, several factors
595 may reduce or increase portability in practice. For example, the impact of individual widgets on
596 final performance is unbalanced because some of them are selected more frequently than others
597 (and thus, their correct recognition is more critical).

598 4.2 Research questions and metrics

600 4.2.1 *RQ1: Feature matching performance.* To achieve the first goal of the study, we formulated the
601 following research question:

602 **RQ1:** How well do feature and template matching algorithms perform when applied to Android
603 GUI widgets?

604 The research question was divided into the following sub-questions:

605 **RQ1.1:** Which widget matching algorithm between Fullscreen and Cropped performs best?

606 **RQ1.2:** Which feature descriptor between SIFT, SURF and AKAZE performs best? How do they
607 compare with template matching approaches?

608 **RQ1.3:** How do feature descriptors and matching algorithms compare in terms of execution
609 time?

610 **RQ1.4:** Which are the main issues for widget matching using feature descriptors?

611 To answer RQ1.1 and RQ1.2, we resorted to precision and recall, standard performance measures
612 for retrieval and classification techniques.

613 We performed an overall performance assessment considering all the widgets extracted from
614 the source and target devices' screen hierarchy. The advantage of this approach is that it enables
615 a comprehensive, large scale and easily automated performance evaluation; however, it does not
616 account for the fact that different types of widgets are selected more or less frequently in test suites
617 and thus have different impacts on the perceived performance of the VGT tool.

618 From the screen hierarchy of each device, we extracted for each widget the bounding box, along
619 with the `content-desc`, `text`, and `resource-id` properties. This information is used to
620 generate the *reference standard* or *ground truth*: for each pair of devices, two graphic components
621 represent the same widget if they have the same id, text, and description. In addition, for each
622 leaf element in the screen hierarchy, we trace the path to its root: if two paths traverse the same
623 containers, preserving the spatial order, they are assigned to the same widget. This constraint allows
624 us to enforce each widget's uniqueness, preserve the relationships between them, and account for
625 cases in which some properties are left empty by the app developer. Another critical aspect to be
626 considered is the timing of the app execution. In particular, the screenshot and dump grabbing
627 must be synchronized and executed after the application is fully rendered to guarantee that the
628 two are correctly aligned.

629 Based on this reference standard, we are able to define which locators are correctly and which
630 are incorrectly matched by the visual algorithms.

631 In particular, given two visual locators, one in the source screen and one in the target screen, we
632 define:

- 633 • *True Positive (TP)*, if the locators correspond to the same widget in the reference standard
634 and are matched by the algorithm;

- *False Positive (FP)*, if the locators do not correspond to the same widget in the reference standard but are matched by the algorithm;
- *False Negative (FN)*, if the locators correspond to the same widget in the reference standard, but they are not matched by the algorithm.

Precision and recall are then calculated as:

$$precision = \frac{\sum TP}{\sum TP + \sum FP};$$

$$recall = \frac{\sum TP}{\sum TP + \sum FN}.$$

To answer RQ1.3, we measured the total time needed for the feature matching algorithms to be applied to the subject images, including feature extraction and matching. In the Fullscreen algorithm, we estimated the processing time as the time to process the entire source and the target screens, plus the additional processing time due to the clustering phase. In the Cropped algorithm, we estimated the time to recover the coordinates for a widget as the sum of the descriptor calculation time and the clustering phase. All calculations prudentially refer to the worst-case scenario in which a single widget is matched on each target screenshot; in the case of multiple widgets to be matched simultaneously, the Fullscreen algorithm's processing time should be divided by the number of widgets to be matched.

4.2.2 *RQ2: Application to GUI testing.* To achieve the second goal of the study, we formulated the following research question:

RQ2: How can feature matching algorithms enhance the portability of GUI test cases in the Android domain?

The research question can be divided into the following sub-questions:

RQ2.1 : How do feature description algorithms compare with state-of-the-art visual testing tools in terms of portability of test scripts to different devices?

RQ2.2 : How do feature description algorithms compare with state-of-the-art visual testing tools in terms of performance?

To answer RQ2.1, we measured the following metrics on the executions of visual test cases on different devices:

- the number of test cases that are executed correctly on different devices;
- the number of correctly identified unique locators;
- the number of correctly performed interactions.

To answer RQ2.2, we measured the total execution time of the scripts, as well as the average time obtained by dividing the former time by the total number of interactions.

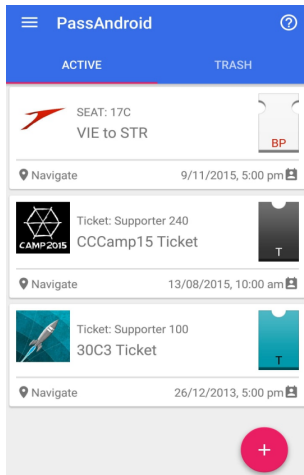
4.3 Experimental subjects and instruments

4.3.1 *Selected applications.* For the first research question, we mined applications from the UpToDown .apk store², a marketplace providing more than 50k free Android apps already used as a source for experimental studies [44][45]. We limited our work to a specific category of Android apps to avoid considering applications with very different graphical appearances, e.g., games or apps with prominent multimedia content. We selected the *Writing and Notes* category, containing mostly utilities to manage and organize text-based content and item lists. We mined with a Python script the entire population of 195 apps belonging to the selected category as of January 2020. We also

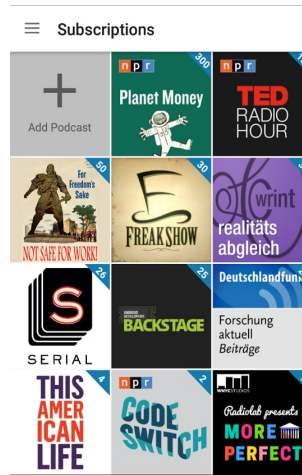
²<https://en.uptodown.com/>

Table 3. Details about the experimental subjects selected for RQ2

	PassAndroid	AntennaPod
GitHub commits	1,697	7,281
GitHub releases	104	98
GitHub stars	542	3.4k
GitHub forks	120	928
PlayStore downloads	1M+	500k+
PlayStore rating	4.3*	4.7*



(a) PassAndroid - Pass Activity



(b) AntennaPod - Subscriptions Activity

Fig. 6. Sample screen captures of the experimental subjects for RQ2

verified whether the apps were compatible with the set of emulated devices that we selected for our analysis. After this verification phase, we came up with a population of 95 valid apps.

For the second research question, we selected two different experimental subjects: PassAndroid [8, 22, 43, 61], a tool for storing and managing different types of tickets through QR codes belonging to the Writing/Notes category of Android apps; AntennaPod [33, 35, 40, 50, 56], an application for listening and managing podcast subscriptions, belonging to the Multimedia/Video category of Android apps. The two apps are popular and long-lived open-source projects on GitHub and are available on the PlayStore and FDroid. We report details about the two apps in Table 3 and sample screen captures in Figure 6.

The test suites were developed by one of the authors of this study with the Appium automation framework. The PassAndroid test suite was partly based on an existing test suite used for a previous study [23]. The author was given no indication about the purpose of the test suite and the designed experiment to avoid possible biases in creating the test scripts. We generated the screen captures and visual locators for each interaction with the GUI by tracing the Layout-based test suite’s execution with proper callbacks.

The characteristics of the test suite are reported in Table 4. The suite consists of 30 different test cases with a varying number of interactions. Not all the interactions of the test suite require a

Table 4. Details about interactions and locators of the test suite developed for the second experimental goal

Property	PassAndroid			AntennaPod		
	Total	Avg.	Median	Total	Avg.	Median
Number of interactions	190	6.3	6	306	10.2	11
Interactions requiring visual locators	170	5.6	5	234	7.8	8
Number of interacted widgets	46	5.3	3	77	7.4	8
Number of different locators	52	5.3	5	77	7.4	8

visual locator (e.g., *PressBack* or *GoHome* are directly executed by the tool through calls to APIs of the ADB and not by interacting with the emulated GUI). The actual number of interacted widgets is markedly lower than the number of interactions performed in the tests, meaning that there are multiple interactions with the same widgets in different tests or even in the same test. It is also worth noting that, in the case of PassAndroid, the number of different visual locators is higher than the number of unique interacted widgets. This fact means that in different test cases or different execution moments, the same widget may have different graphical appearances.

4.3.2 Feature matching implementation details. We used the latest releases (at the time of the experiment, in June 2020) of the Python version of OpenCV and RANSAC libraries (3.4.2.17).

In this work, we compare three scale, rotation and translation invariant feature matching techniques: SIFT [39], SURF [14] and AKAZE [29, 58]. They represent classes of descriptors with different trade-offs in terms of accuracy, memory occupation, and execution time [55]. AKAZE was already successfully used in VGT tools [58]. It belongs to the class of binary descriptors (like ORB and BRISK) and offers a reduction of the computational burden. To perform the actual matching, we use brute force matching, i.e., each keypoint is associated with the closest one in the feature space using as a metric the Euclidean distance for SIFT and SURF and the Hamming distance for AKAZE.

As for the RANSAC, we used a recent variant named Graph-Cut RANSAC [13], which is faster and more accurate than standard RANSAC. Graph-Cut RANSAC is applied with the following parameters: the smallest number of data points to evaluate model parameters is set to 10, the maximum number of iterations to 1000, and the threshold value (to determine which data points are fit by the model) is set to 100.

4.3.3 State-of-the-art VGT tools. As state-of-the-art tools to be included in the comparison, we selected EyeAutomate [6], release 2.2, and SikuliX [62], release 1.1.2, since they are the most cited in empirical studies on visual testing. We have used the tools by leveraging the provided APIs in Java.

To achieve a more systematic comparison with state-of-the-art tools for RQ1, we re-implemented the matching algorithm employed by the open-source software SikuliX using the same library (OpenCV) and settings employed in the tool. This choice made it possible to extract the raw matching performance of SikuliX on a widget-by-widget basis, exploiting the experimental setup and script developed for RQ1. This approach could not be used for EyeAutomate since the latter is not open source and leverages a proprietary algorithm on which it was not possible to apply reverse engineering.

4.3.4 Android virtual devices. As our emulated Android devices set, we leveraged the 14 default devices in the Android AVD Manager. The properties of the considered devices are reported in Table 5. All the devices used Android API 25 (version 7.11) and mounted x86 system images. The emulated devices were not hardware-accelerated, had device frame and keyboard inputs enabled,

Table 5. List of emulated devices considered for the research

Device model	Screen size (pixels)	Resolution
Nexus 5	1080 × 1920	xxhdpi
<i>Pixel 3</i>	1080 × 2160	440dpi
<i>Pixel 2</i>	1080 × 1920	420dpi
<i>Nexus 5X</i>	1080 × 1920	420dpi
<i>Nexus 6P</i>	1440 × 2560	560dpi
<i>Nexus 6</i>	1440 × 2560	560dpi
Nexus S	480 × 800	hdpi
<i>Pixel 3a</i>	1080 × 2220	440 dpi
<i>Pixel 3a XL</i>	1080 × 2160	402 dpi
<i>Pixel</i>	1080 × 1920	420 dpi
<i>Nexus 4</i>	768 × 1280	xhdpi
Pixel 3 XL	1440 × 2960	560 dpi
<i>Pixel 2 XL</i>	1440 × 2880	560 dpi
<i>Pixel XL</i>	1440 × 2560	560 dpi

while all the animations were disabled to avoid errors in transitions between Activities. In the table, we reported in bold the devices that we used as sources for the experiments. As it can be seen, we selected three devices with very different resolutions and pixel densities.

4.3.5 Experimental setup. All the experiments were performed on a desktop PC with an Intel i7-4770 running at 3.40GHz clock, with 8GB RAM and Ubuntu 20.04 LTS 64-bit as OS.

4.4 Procedure

4.4.1 RQ 1. We designed and implemented an automatic procedure for performance assessment across our app's database; this procedure allows extensive validation with minimal human effort.

We evaluated the feature matching algorithms on all the widgets shown in the main Activities for each pair of source and target devices. We only used these Activities to avoid the need to navigate the different screens of the apps. We leveraged the Android emulator's debugging capabilities to retrieve from all the devices detailed information about these Activities; to this aim, we used the *dump files* containing all properties of the current visualized widgets.

The dump files' information can emulate the cropping and bounding box drawing operations that the tester would perform to prepare the visual test suite. It should be stressed that the information contained in the dump files is not used by the matching algorithm (which relies purely on the visual content) but is merely exploited to automate the assessment procedure.

In our experiments, we extracted all possible widgets from the source device screen and attempted to find the corresponding visual locators in the target device screen with different algorithms and descriptors. We repeated the procedure on 3×13 pairs of devices, using one of the three selected devices as source and the remaining 13 devices as the potential target. Recall and precision were separately computed for each app, then their distribution was calculated over the entire database grouping by the statistical factors of interest.

Like in a real test case, we assume to perform matching based only on the visible components. Due to fragmentation, some components may be rendered on the source device but not on the target device (Figure 1). Since our approach is purely visual, such locators are not included in the ground truth, and any accidental matching would be counted as an FP. Finally, it is worth underlining that a VGT tool would fail if the missing component is included in a test suite. This entails that even

834 a perfect recall does not guarantee, in principle, perfect test portability. This aspect is taken into
835 account in RQ2 and the analysis of the generated dataset.

836
837 4.4.2 *RQ 2.* We collected the screen captures and the cropped widget locators for all interactions
838 in the test suite on the three devices we used as sources. Then, we executed the test cases on all 14
839 devices of the set and measured the number of correctly identified locators, correctly performed
840 interactions, and completely executable test cases. A test case is considered completely executable
841 if a given feature matching algorithm correctly identifies all locators used in it. We also considered
842 the situation in which the source and target devices coincide since the matching algorithms may
843 yield wrong coordinate pairs even if the test script is applied on the same device where the locators
844 were captured (e.g., in the case of multiple widgets with the same appearance). The application
845 of a test script on the original device is a common scenario since VGT techniques can be used for
846 regression testing on new releases of the application.

847 We ran all test cases by embedding the feature matching algorithms (which provide as output
848 the coordinate pairs of the identified widgets) in scripts that executed the found coordinates'
849 interactions by using the Appium library. We verified the correctness of the resulting coordinates
850 by checking the dump files obtained after each interaction.

851 We then used EyeAutomate and Sikuli with the cropped screen captures to replicate the test
852 executions. This time we used only the Cropped captures since the pixel-per-pixel comparisons
853 used by the tools would mostly lead to failures in recognizing fullscreen captures even in the
854 presence of minor changes in device screen sizes.

855 Visual test scripts were always executed on a solid black background to minimize other visual
856 elements' interference. No other computationally intensive program was run concurrently to avoid
857 external influences on execution times.

858
859 4.4.3 *Statistical analysis.* A multi-way factorial Permutational Analysis of Variance (PERM-ANOVA)
860 [9] was conducted to examine the main effects of matching algorithm, descriptor, target, and source
861 device, as well as the interaction effects between target and source devices on each metric defined
862 for research questions RQ1 and RQ2. PERM-ANOVA is a non-parametric multivariate statistical test.
863 The null hypothesis tested by PERM-ANOVA is that the centroids of the partitions or groups defined
864 by a similarity metric (e.g., the Euclidean distance) are equivalent for all groups. Exact p-values are
865 obtained by calculating the test statistics' value for all (or a large random subset) of permutations of
866 the observations across different groups. In particular, the proportion of the values of the statistics
867 under different permutations (i.e., random re-allocation of individual samples to different groups)
868 that are equal to or higher than the observed value. If the null hypothesis was true, any observed
869 differences would be similar to those obtained under permutation. PERM-ANOVA only requires
870 exchangeability and does not make any assumption on the sample distribution, accommodating
871 severely non-normal variables, contain a large number of zeros, or are ordinal or qualitative in
872 nature [9].

873 For RQ1, the matching algorithm and feature descriptors were modeled as separate fixed factors,
874 along with their interaction. For RQ2, we considered the overall technique as a fixed factor, which
875 could be either one of the combinations of the descriptor and matching algorithm (e.g., SIFT -
876 Fullscreen) or one of the state-of-the-art VGT tools. Each app was modeled as a different subject
877 with repeated measures. Post-hoc comparison between the different combinations of matching
878 algorithms and feature descriptors was performed using distribution-free pairwise two-sample
879 permutation tests [41] applying Bonferroni correction. We also measured the effect size for any
880 pair of groups of observations by using Cliff's delta formula. Statistical analysis was performed in
881 R, and the *effsize* package was adopted for effect size computation [57].

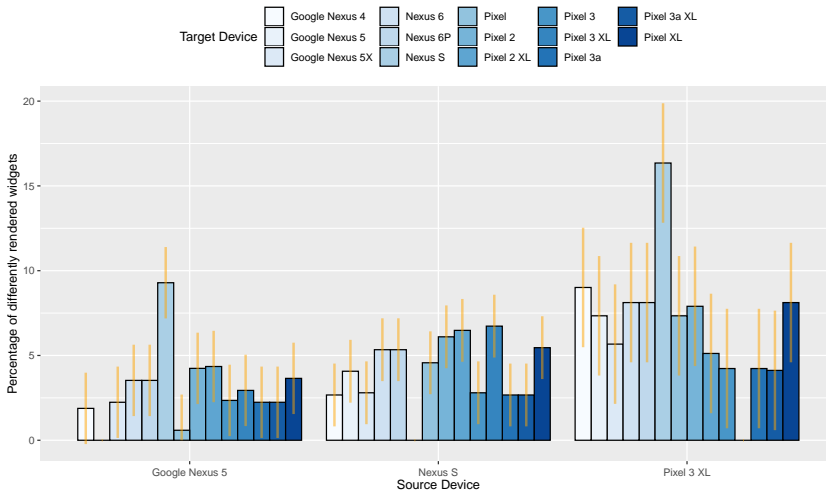


Fig. 7. Percentage of widgets that are present in the source device but not in the target device, based on the comparison of the XML trees in the dump files. The percentage is calculated separately for each source and target device pair and is higher for device pairs with very different screen sizes. Yellow lines report the intervals of confidence calculated on a binomial distribution.

The raw data and markdown scripts have been made available on a GitHub repository³. Null hypotheses, detailed results for the post-hoc comparison and values of effect size are reported in the supplementary material of the present manuscript⁴.

5 RESULTS

5.1 Dataset characteristics

In this section, we describe the characteristics of the subjects collected for the RQ1 study. In the population of 95 apps, a total of 787, 850, and 900 unique widgets were identified for the Nexus S, Nexus 5, and Pixel 3 XL devices, respectively.

As mentioned in Section 2.3, some widgets could not be localized in both the source and target devices. A number of possible causes were identified for this phenomenon. Most commonly, Android performs rescaling and resizing operations to optimize the User Experience, adapting the interface layout to the screen size, which in turn changes the position, dimension, and resolution of the widgets. Some widgets may be grouped together to reduce visual clutter on small devices. In fact, the average number of individual widgets per app is proportional to the screen size of the emulated device, ranging from 8.11 (Nexus S) to 9.28 widgets/app (Pixel 3). Less frequently, discrepancies between the source and target devices arise due to banners incorporated in the app’s layout, which are selected randomly and remain on screen with an automatic refresh time.

The above-mentioned differences may constitute an intrinsic limitation to the performance of visual matching algorithms, at least in the current Record and Replay scenario, which assumes that the interactions recorded on the source device, and their visual locators, can be found and replicated on the target device. To estimate the order of magnitude and potential impact of this issue, we

³https://github.com/SoftengPoliTo/image_matching_study

⁴https://figshare.com/articles/online_resource/Feature_Matching-based_Approaches_to_Improve_the_Robustness_of_Android_Visual_GUI_Testing_Supplementary_material/14912292

Table 6. Average values (and std. deviation) of precision, recall and execution time, grouped by technique.

	Recall	Precision	Execution time (s)
SIKULI - Cropped	0.38 (0.33)	0.46 (0.37)	0.41 (0.49)
AKAZE - Cropped	0.71 (0.31)	0.84 (0.26)	0.69 (0.71)
SIFT - Cropped	0.95 (0.12)	0.91 (0.17)	1.0 (0.51)
SURF - Cropped	0.91 (0.18)	0.86 (0.21)	0.99 (1)
AKAZE - Fullscreen	0.87 (0.20)	0.92 (0.19)	1.28 (0.7)
SIFT - Fullscreen	0.88 (0.19)	0.93 (0.18)	1.48 (0.63)
SURF - Fullscreen	0.89 (0.18)	0.92 (0.18)	1.31 (1.06)

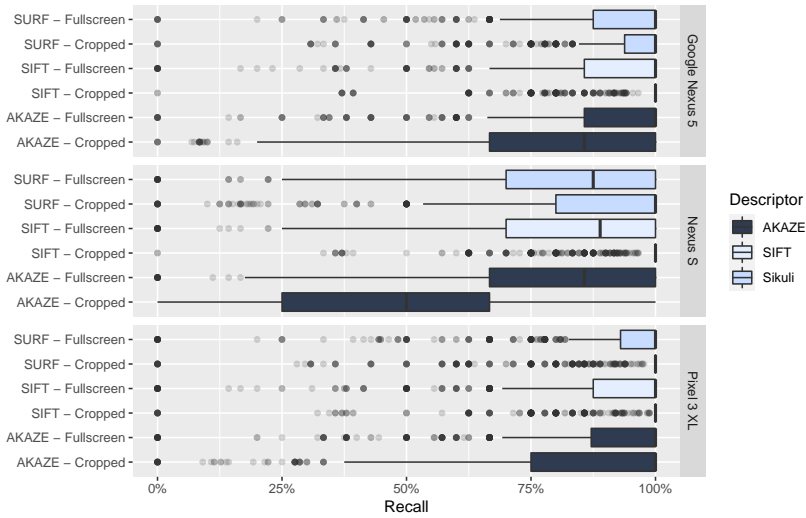


Fig. 8. Distribution of the recall over the set of apps and target devices, per source device and technique. Recall is reported for each combination of descriptor and matching algorithm, Fullscreen (blue) vs. Cropped (light blue).

calculated for each device pair the percentage of widgets (mean and confidence interval) that were present in the ground truth of the source device but not in the ground truth of the target device, and reported the distribution in Figure 7. It should be noticed that this percentage is independent of the specific algorithm or visual testing tool and depends solely on the combination of source and target devices. The percentage of widgets that cannot be correctly located is in most cases well below 10%, except for the two devices with the largest difference in resolution and screen size (Nexus S and Pixel 3 XL).

5.2 RQ1: Feature matching performance

Table 6 shows the average and standard deviation values for the metrics selected for RQ1, namely recall, precision, and execution time.

The distribution of the recall and precision is reported in Figures 8 and 9. Recall and precision are calculated for each app and for each target device and then grouped by source device and technique. Overall, there is a statistically significant difference across all groups for both precision (p -values $< 2e-16$) and recall (p -values $< 2e-16$). In our results, precision and recall exhibit similar behavior: since a given source widget on the source device can be associated only to a single target widget, a

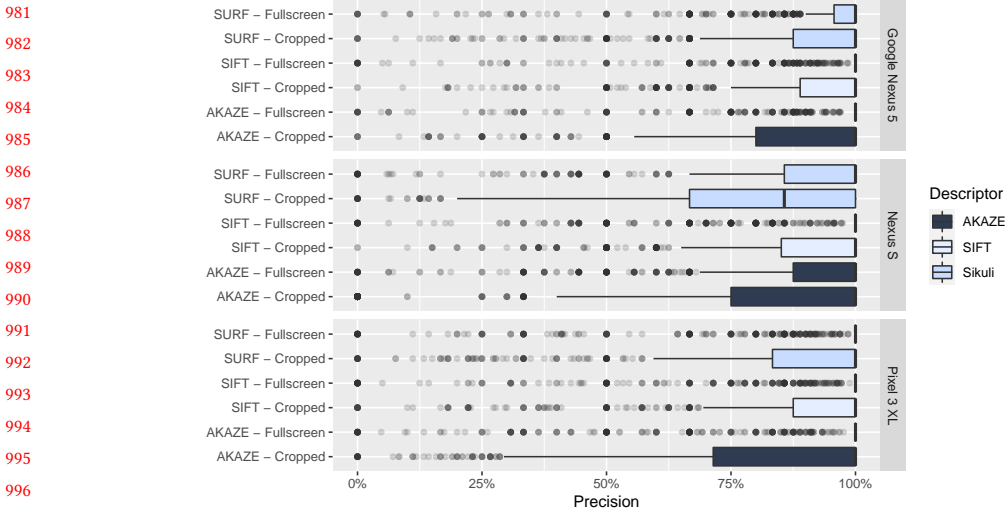


Fig. 9. Distribution of the precision over the set of apps and target devices, per source device and technique. Recall is reported for each combination of descriptor and matching algorithm, Fullscreen (dark blue) vs. Cropped (light blue).

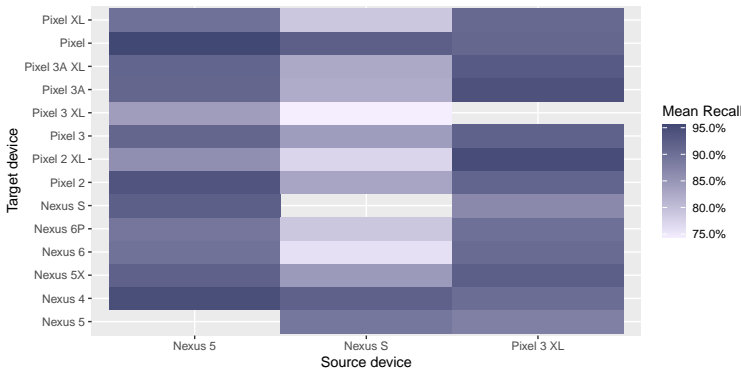


Fig. 10. Average recall of the SIFT - Fullscreen combination, with varying source and target devices.

correct matching would increase both precision and recall, whereas an incorrect matching would affect both. We thus report for brevity a detailed analysis only for recall.

We observed a significant effect of matching algorithm, descriptor, and their interaction on both precision and recall ($p < 1e-10$). At post-hoc analysis, the Fullscreen technique achieved higher recall than the Cropped technique for the AKAZE ($p < 1e-10$), SIFT ($p = 2e-07$), and SURF descriptors ($p < 1e-10$). For the Cropped technique, SIFT and SURF outperformed both the AKAZE descriptor ($p < 1e-10$) and Sikuli template matching ($p = 0.0$); SIFT also achieved higher recall than SURF ($p = 0.0$). For the Fullscreen technique, SURF slightly outperformed both AKAZE ($p = 3.5e-05$) and SIFT ($p = 2e-13$), whereas differences between SIFT and SURF were not statistically significant ($p = 0.09$). Sikuli can only operate in the Cropped modality, hence post-hoc comparison with the Fullscreen algorithms was not attempted. The worst performing technique was Sikuli, followed by

1030 the Cropped algorithm with the AKAZE descriptor. In general, the Fullscreen algorithm appeared
1031 more robust to the choice of the descriptor.

1032 The source and target devices had a significant effect on the recall ($p < 1e-10$) and, as expected,
1033 a significant interaction ($p < 1-10$). The best performance was obtained when using the Pixel 3
1034 XL as a source (the device with a larger screen size and resolution), whereas starting from the
1035 smallest device (Nexus S) as the source device, generally poorer results were observed. This is
1036 further illustrated by the color map in Figure 10, reporting the recall for source and target device
1037 pairs for the best performing technique (i.e., SIFT - Fullscreen).

1038 Both algorithms were, on average, quite successful in locating widgets. Across all tested source-
1039 target device pairs and apps, roughly 60% of the cases achieved a perfect recall of 100% (6431/10997
1040 for the Fullscreen and 6879/10997 for the Cropped algorithm), meaning that all source widget
1041 locators could be correctly matched on the target device, whereas roughly 75% of them (8208/10997
1042 and 7919/10997) achieved a recall higher than 80%. Therefore, the distribution is highly skewed on
1043 the left side, which explains a large number of outliers in the boxplots reported in Figures 8 and 9.
1044 Nonetheless, in a small number of cases, 1% (185/10997) and 4% (480/10997) for the Fullscreen and
1045 Cropped algorithms respectively, the recall was below 25%, i.e., most of the visual locators could
1046 not be identified on the target device.

1047 At visual inspection, for the Fullscreen algorithm, such outliers appear to be mostly associated
1048 with widgets (including buttons) that display long text strings. Widgets that include text are
1049 processed as any other widget and, usually provide many robust keypoints, thus facilitating many-
1050 to-many feature matching. However, since the latter process is entirely visual and does not take
1051 into account the text semantics, the same letters can be matched in different words, leading in a
1052 few cases to a high number of FPs. RANSAC becomes less effective in eliminating outliers as the
1053 number of FPs increases.

1054 The Cropped algorithm may fail when the keypoints in the source visual locator are matched
1055 with multiple target widgets. This can be due to repeated or similar widgets (see the example
1056 in Figure 4) or, like for the Fullscreen algorithm, to the presence of long text strings. In these
1057 settings, RANSAC does not have enough information to select the correct matching based on spatial
1058 consistency, which would require considering all the widgets in the source and target devices
1059 simultaneously. Thus, the results of the final clustering may be wrong.

1060 Finally, the execution time is reported in Figure 11. The average processing time is higher for
1061 the Fullscreen algorithm (due to the need to calculate all the keypoints in the source image) and
1062 increases with the size of the device screen. A few outliers can be observed mostly due to apps with
1063 long text strings, which increase the number of matches. As expected from the literature, AKAZE
1064 is faster than SURF, whereas SIFT is the slowest descriptor.

1065 5.3 RQ2: Application to GUI testing

1066 Table 7 reports the average and standard deviation values for the metrics measured to answer
1067 RQ2. The average values include the measurements on both the applications considered in the
1068 experimental procedure. Feature matching-based algorithms outperformed the state-of-the-art
1069 VGT tools in all the aspects related to RQ2.1. EyeAutomate and Sikuli showed lower portability for
1070 the three evaluated metrics with all the source devices. Table 8 and table 8 report the average and
1071 standard deviation values for the individual SUTs considered for the experiment.

1072 Regarding the executed interactions, feature matching algorithms guaranteed a percentage of
1073 correctly executed interactions that ranged from 72% (for AKAZE - Cropped) to 95% (for SURF -
1074 Cropped), significantly higher than state-of-the-art tools, between 43% (Sikuli) and 59% (EyeAuto-
1075 mate). Hence, for the considered combinations of apps and emulated devices, our methodology
1076 increased the percentage of correctly executed instructions by at least 30% compared to the best
1077

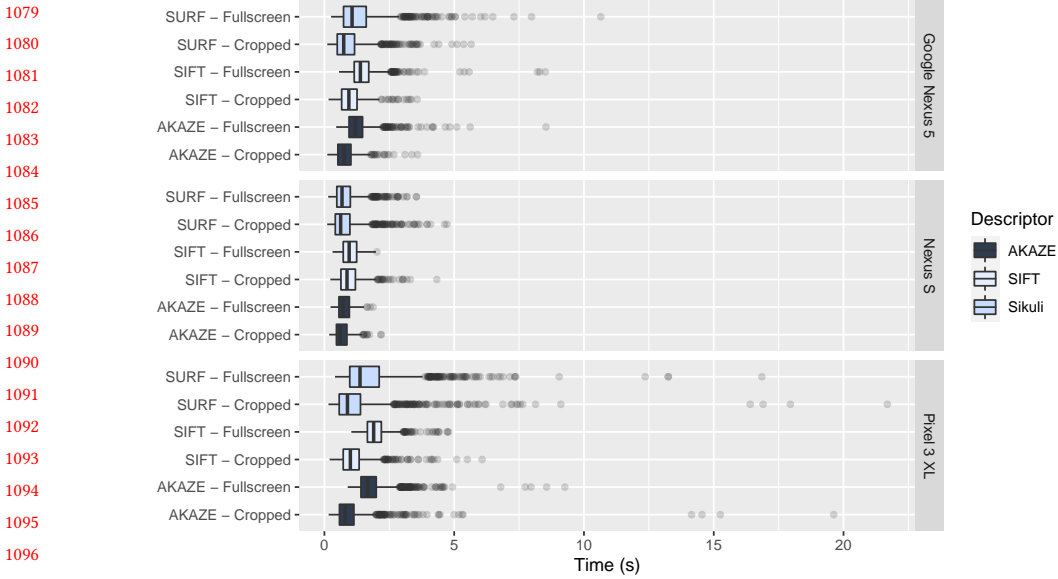


Fig. 11. Distribution of the execution time (in seconds) over the set of apps and target devices, per source device and technique. Recall is reported for each combination of descriptor and matching algorithm, Fullscreen (blue) vs. Cropped (light blue).

Table 7. Average values (and std. deviation) of the percentage of executed interactions, found locators, passing tests and time per interaction over all executions of the test suites, grouped by technique

	Proportion of correctly executed interactions	Proportion of found locators	Proportion of correctly executed tests	Time per interaction
EyeAutomate	0.59 (0.25)	0.60 (0.23)	0.17 (0.22)	548 (313)
Sikuli	0.43 (0.24)	0.31 (0.26)	0.11 (0.27)	373 (134)
AKAZE - Cropped	0.72 (0.18)	0.67 (0.22)	0.24 (0.29)	1238 (458)
SIFT - Cropped	0.89 (0.09)	0.92 (0.08)	0.50 (0.37)	2041 (904)
SURF - Cropped	0.82 (0.16)	0.81 (0.15)	0.43 (0.32)	1925 (1057)
AKAZE - Fullscreen	0.94 (0.06)	0.94 (0.06)	0.76 (0.24)	1927 (1261)
SIFT - Fullscreen	0.95 (0.06)	0.96 (0.04)	0.77 (0.25)	2712 (1836)
SURF - Fullscreen	0.95 (0.06)	0.95 (0.04)	0.75 (0.22)	2737 (2044)

performing state-of-the-art VGT tool (EyeAutomate). We observed a significant effect on the percentage of found locators of the matching algorithm, descriptor, source, and target device, and the combination of source and target devices (with p -values $< 10e-16$). At post-hoc analysis, we verified that the SURF - Fullscreen technique outperformed in a statistically significant way all the Cropped matching algorithms and the VGT tools in terms of correctly executed interactions (with p -values ranging from $4.64e-07$ for the comparison with SIFT - Cropped to $2.89e-26$ for the comparison with Sikuli).

The boxplot in Figure 12 reports the percentage of found locators for each source and target pair, grouped by technique. Sikuli was the worst option (31% of found unique locators), whereas SIFT - Fullscreen proved to be the best one (96%), with 36% more locators found than the best performing state-of-the-art VGT tool analyzed (EyeAutomate). We observed a significant effect on the percentage of found locators of the matching algorithm, descriptor, source and target device,

Table 8. Average values (and std. deviation) of the percentage of executed interactions, found locators, passing tests and time per interaction over all executions of the test suites, for the PassAndroid SUT

	Proportion of correctly executed interactions	Proportion of found locators	Proportion of correctly executed tests	Time per interaction
EyeAutomate	0.64 (0.24)	0.59 (0.19)	0.26 (0.24)	711 (399)
Sikuli	0.48 (0.24)	0.37 (0.26)	0.15 (0.29)	418 (139)
AKAZE - Cropped	0.83 (0.12)	0.78 (0.16)	0.37 (0.34)	1268 (487)
SIFT - Cropped	0.93 (0.06)	0.95 (0.03)	0.73 (0.24)	1476 (586)
SURF - Cropped	0.91 (0.06)	0.84 (0.11)	0.62 (0.22)	1160 (440)
AKAZE - Fullscreen	0.94 (0.07)	0.93 (0.06)	0.78 (0.21)	1144 (388)
SIFT - Fullscreen	0.93 (0.07)	0.96 (0.04)	0.74 (0.30)	1287 (431)
SURF - Fullscreen	0.96 (0.03)	0.94 (0.04)	0.85 (0.12)	1011 (313)

Table 9. Average values (and std. deviation) of the percentage of executed interactions, found locators, passing tests and time per interaction over all executions of the test suites, for the AntennaPod SUT

	Proportion of correctly executed interactions	Proportion of found locators	Proportion of correctly executed tests	Time per interaction
EyeAutomate	0.53 (0.25)	0.61 (0.26)	0.90 (0.17)	385 (169)
Sikuli	0.38 (0.23)	0.26 (0.26)	0.08 (0.26)	328 (114)
AKAZE - Cropped	0.63 (0.21)	0.56 (0.21)	0.11 (0.14)	1225 (445)
SIFT - Cropped	0.84 (0.09)	0.89 (0.10)	0.28 (0.32)	2610 (816)
SURF - Cropped	0.74 (0.18)	0.77 (0.17)	0.25 (0.29)	2695 (943)
AKAZE - Fullscreen	0.95 (0.06)	0.95 (0.05)	0.73 (0.26)	2711 (1345)
SIFT - Fullscreen	0.97 (0.04)	0.96 (0.04)	0.80 (0.19)	4137 (1573)
SURF - Fullscreen	0.93 (0.07)	0.95 (0.04)	0.65 (0.26)	4399 (1490)

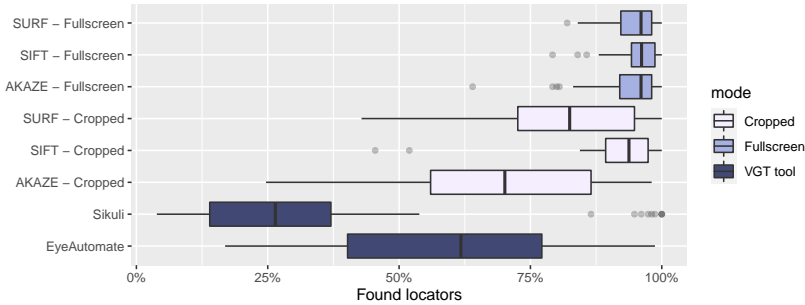


Fig. 12. Distribution of the percentage of found locators for each source and target device pair, grouped by technique. Available tools (dark blue), Fullscreen (blue), Cropped (light blue).

and the combination of source and device (with p -values $< 10e-16$). When using the Fullscreen algorithm, the percentage of found locators did not differ significantly for each descriptor. The heatmap in Figure 13 reports the percentage of found locators per source and target device pair for the best performing technique (i.e., SIFT - Fullscreen).

The plot shows lower percentages of found locators when the Nexus S was selected as either the source or target device. This outcome was likely due to the small size of the device screen (480×800 pixels). The average percentage of passing test cases was lower than that of executed instructions and found locators. This result was mainly due to the fact that the same unique locator can be used multiple times in different tests: e.g., in PassAndroid, most of the test cases involve a click on the *floating action button*, which systematically leads to a FP when the AKAZE - Cropped

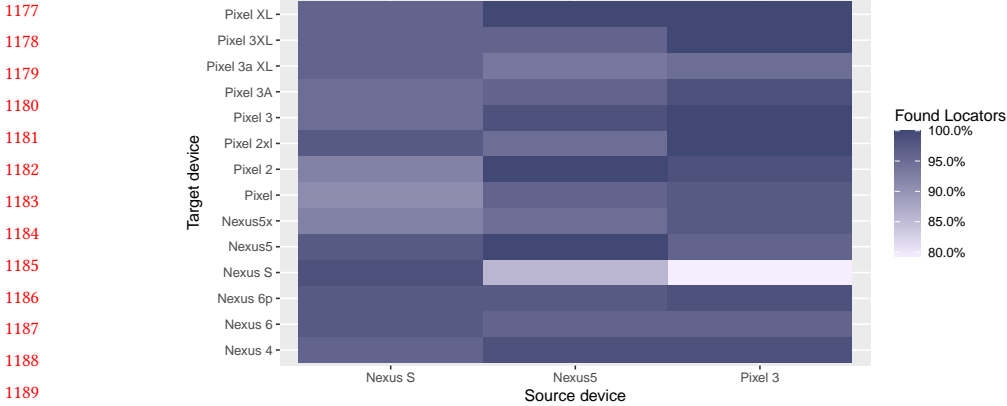


Fig. 13. Percentage of found locators during the test suite execution with the SIFT - Fullscreen technique, with varying source and target devices.

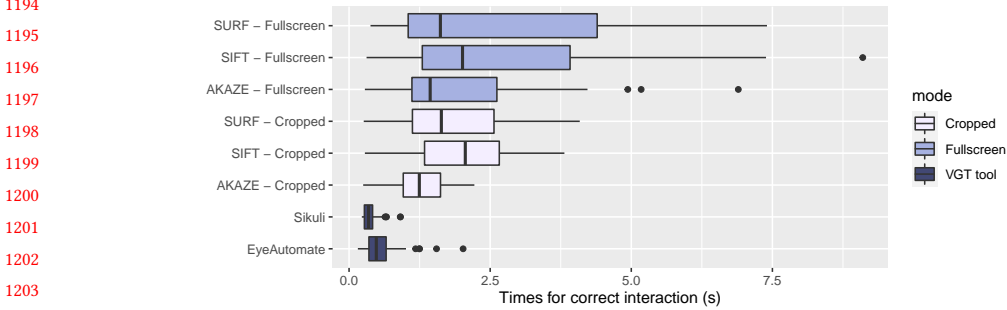


Fig. 14. Distribution of the average time per correct interaction for each source and target device pair, grouped by technique.

technique is used; in AntennaPod, many test cases involve the execution of a click on the Android menu button, which in most cases is not recognized by EyeAutomate. However, the percentages of passing tests were very high for the Fullscreen algorithm, regardless of the descriptor used, with 77% for SIFT - Fullscreen. Only 17% and 11% of the test cases were successfully executed for state-of-the-art VGT tools EyeAutomate and Sikuli, respectively. We observed a significant effect of matching algorithm, descriptor, source, and target device on the percentage of passing tests, as well as of the combination of source and target (with p -values $< 10e-16$). At post-hoc analysis, the SURF - Fullscreen configuration outperformed all other techniques in a statistically significant way, except for AKAZE - Fullscreen ($p = 0.06$). For feature matching algorithms, with post-hoc tests we measured statistically significant differences except for AKAZE - Fullscreen, SIFT - Fullscreen, and SURF - Fullscreen.

Figure 14 reports the distribution of the average time needed by the VGT tools to perform a correct interaction (i.e., time to identify a widget plus time to execute interaction), grouped by technique. VGT tools exhibited a lower execution time per interaction, with an average time of 373 milliseconds for Sikuli and 548 milliseconds for EyeAutomate. The fastest feature matching technique was AKAZE - Cropped (1.24 seconds per interaction), whereas SIFT - Fullscreen was the

slowest (2.74 seconds per interaction). We observed a significant effect on the percentage of passing tests of matching algorithm, descriptor, source, and target device (with p -values $< 10e-16$), whereas no significant interaction was found between the time to find a locator and the combination of source and target devices ($p = 0.99$).

As a final analysis, we observed the effect of the selected app (in our case, PassAndroid vs. AntennaPod) on the controlled variables. From the average and median values reported in Table 8 and 9 we can see that the SIFT - Fullscreen algorithm performed best for all metrics with the AntennaPod SUT. Conversely, the SURF - Fullscreen was the best algorithm for the PassAndroid SUT regarding the percentage of executed interactions and the proportion of correctly executed tests, whereas SIFT - Fullscreen was still the best algorithm in terms of found locators. Sikuli was the best performing algorithm in terms of time per interaction for both SUTs. The impact on the results was more marked when the Cropped algorithm, rather than Fullscreen algorithm, was used. This result can be reasonably justified with the assumption that the performance of the Cropped algorithm strictly depends more on the nature of the individual SUT, since the results can be strongly impacted by the different arrangement of widgets in the layouts.

We observed a statistically significant effect on the percentage of found interactions, correct tests, and average time per interaction (p -value $< 2.2e-16$). The selected app had no significant effect on the percentage of correctly found locators (p -value = 0.094). This result strongly suggests that the ability to find individual locators of the feature matching algorithms is unrelated to the AUT. The percentage of passing tests and executed interactions, on the contrary, strongly depends on the way the individual test suite has been defined (i.e., the number of locators used in each test case and the repetition of locators in different test cases). The time to execute the feature matching algorithms is also expected to be correlated to the AUT since it depends on the total number of objects on the screen that have to be examined.

6 DISCUSSION

The performance of VGT tools strongly depends on the performance and quality of the underlying image analysis technique. By systematically exploring different combinations of feature matching algorithms and feature descriptors, we observed that the portability of test suites across mobile devices could be substantially improved over state-of-the-art tools.

For RQ1, we compared different algorithms from a purely visual matching perspective, investigating how often, on average, is it possible to correctly match any given widget from a source device through its relative visual locator on a target device.

Overall, the Fullscreen algorithm proved more robust and precise in locating widgets. By determining the optimal matching for all widgets simultaneously, it can solve complex cases (e.g., repeated widgets) where the Cropped algorithm is likely to fail. The difference is striking for the AKAZE descriptor, which also achieves the lowest overall performance. This result is consistent with the literature indicating that AKAZE is more computationally efficient but less robust to downscaling than SIFT and SURF [55].

For RQ2, we compared how well the different algorithms performed, in terms of robustness of the test cases based on them.

The number of correctly found locators in RQ2 is consistent with the recall in RQ1, and the same trend emerges in both the analyses with respect to both matching algorithm and descriptor. Among the configurations of feature descriptors and algorithms, SURF - Fullscreen is the best solution in terms of the percentage of correct interactions and of test suites executed. SIFT - Fullscreen is the best combination in terms of percentage of found locators. Based on the combined results of RQ1 and RQ2, both SIFT and SURF emerge as viable options, with SIFT achieving slightly higher performance and SURF being slightly faster.

1275 The procedure followed in RQ1 can be easily replicated by researchers willing to evaluate other
1276 matching algorithms that can be employed in VGT of Android applications. The methodology sys-
1277 tematically and automatically compares the algorithms across applications and devices, bypassing
1278 the need to define test cases manually. The procedure, however, does not consider the varying
1279 relevance of different types of widgets in real-world test suites, in which the performance of the
1280 matching algorithms with individual widgets can influence the result of the execution of several
1281 test cases. In our second experiment, we measured an average percentage of passed tests that were
1282 much lower than found locators (with a difference between -20% and -30%), primarily due to a
1283 single mismatched locator. This result suggests that further research work is needed to assess and
1284 improve the performance selectively on the most relevant widgets for real test cases.

1285 When the source is equal to the target device, the EyeAutomate and Sikuli tools showcased very
1286 high percentages of found locators (and therefore of correct interactions and passed tests). However,
1287 they showed low portability across devices, with EyeAutomate consistently outperforming Sikuli
1288 (in accordance with previous studies [21]). The proposed feature matching techniques have higher
1289 overall portability, increasing the percentage of found locators and passed tests by at least 30% with
1290 respect to state-of-the-art VGT tools.

1291 We postulate that the performance gap arises from the combination of two factors: the matching
1292 strategy and the feature extraction. State-of-the-art VGT tools like EyeAutomate and Sikuli are all
1293 based on the less performing Cropped algorithm, whereas our results highlight that the task is best
1294 tackled as a many-to-many matching problem where all widgets are simultaneously optimized.
1295 Secondly, SIFT and SURF provide more robust features than pixel-level template matching. In
1296 particular, their invariance to scale is of paramount importance when covering a wide range of
1297 screen sizes.

1298 These results suggest that state-of-the-art VGT test suites may provide sufficient robustness
1299 when used only on a single device, e.g., for regression testing purposes. Conversely, when the
1300 portability to different devices is important, our algorithms based on feature description algorithms
1301 are preferable. A further advantage of the proposed matching algorithms is that they do not require
1302 scaling of the screen captures to the size of the AVD as exactly rendered on the host screen, which
1303 on the contrary, is needed when using Sikuli and EyeAutomate.

1304 In terms of execution time, the difference between the Fullscreen and Cropped algorithms is
1305 almost negligible. However, Sikuli and EyeAutomate are faster than the feature matching algorithms,
1306 with Sikuli being the fastest (the decrease of average time per interaction ranges from 70% with
1307 respect to AKAZE - Cropped, to 85% with respect to SURF - Fullscreen). We speculate that the
1308 difference is due to the more effective but more computationally expensive descriptors and to the
1309 search strategy. EyeAutomate starts searching from the last location (or the upper left corner, by
1310 default) [1], whereas our proposed techniques take into account (and compute the features for) all
1311 possible locations within the image. This choice has obvious advantages in the case of multiple
1312 similar widgets, increasing the accuracy, but it is paid for in terms of execution time. We did not
1313 specifically attempt to optimize the feature matching algorithms and their implementation for
1314 execution time, leaving this aspect to future work.

1315 The additional time needed can be a limitation when the locators have to be found in highly
1316 dynamic GUIs, where the widgets and images can change very rapidly (e.g., in games). However,
1317 the feature matching algorithm we propose is commonly used for photos, so it may provide higher
1318 precision when used for complex GUIs than Sikuli and EyeAutomate. Additional comparisons with
1319 graphically-intensive apps should be performed as future work to investigate this aspect.

1320 By performing a large-scale, automatic validation across multiple applications and devices, we
1321 found some limitations to matching algorithms based solely on visual features. The layout of
1322 Android applications is optimized based on the screen size by, e.g., resizing, rescaling, and grouping

1323

1324 widgets, exploiting vertical and horizontal scrolling, etc. In a small but not negligible percentage
1325 of cases, widgets present in the source device cannot be located in the target device, especially
1326 if the difference in screen size is large. In addition, some types of widgets and components (e.g.,
1327 those with long text strings) are more prone to matching errors. These issues may affect a small
1328 number of test suites and hence may remain undetected following the assessment procedure in RQ2.
1329 Many of these issues could be tackled more effectively by integrating feature matching algorithms
1330 with semantic image interpretation capabilities, identifying the type, content, and function of each
1331 widget, and, if needed, modifying the matching strategy or even inserting additional operations
1332 (e.g., scrolling) in the test suite.

1333 The proposed algorithms do not distinguish between text-based and image-based widgets and
1334 do not explicitly seek to identify or interpret the text. For feature-based matching, this operation is
1335 not necessary as the text provides many robust keypoints, which is generally beneficial in terms of
1336 performance but may occasionally create robustness issues for a minority of apps with very long
1337 text strings. In the case of widget detection, a mixed approach was found beneficial over a purely
1338 visual approach [20]. Similar strategies may be integrated in the future by slightly modifying the
1339 feature extraction.

1340 Regardless of the algorithm or tool employed, the selection of the source and target devices had
1341 a statistically significant impact on all performance measures. Very small devices, like Nexus S,
1342 yield significantly worse performance when used as either source or target devices. This fact has
1343 practical implications for developers and testers, who need to consider the range of devices they
1344 wish to support carefully. When the range of target devices is wide, selecting a source device with
1345 medium to large screen size will increase portability and decrease the effort needed to maintain the
1346 test suite.

1347 7 THREATS TO VALIDITY

1349 *External validity threats.* For the first experiment, we considered a single category of apps mined
1350 from a single database. Although this decision sets a realistic context for a high number of Android
1351 apps, it does not ensure that the results that we described are applicable to any type of Android app.
1352 Further studies are needed in games and other highly dynamic GUIs, which require both highly
1353 accurate and fast visual matching algorithms.

1354 To better characterize the properties and generalizability of our results, we have studied the
1355 distribution of different types of widgets in our dataset (additional details are provided in the
1356 Supplementary Material). We further compared this distribution with that of the RICO dataset
1357 [26], which is based on a larger sample of roughly 9300 apps. Roughly 50% of the widgets (40%
1358 in the RICO population) are TextView components. The most frequent widgets include Button
1359 (13%), ImageView (11%), View (6%), and ImageButton (6%). All other categories constitute overall
1360 11% of the total number of widgets. The distribution is qualitatively comparable between our
1361 dataset and the entire RICO population, with a higher prevalence of TextView, EditText, and Button
1362 components in our dataset and a higher prevalence of View and ImageView in the RICO population.
1363 From this analysis, we conclude that any approach designed to solve the VGT issues, especially in
1364 the mobile domain, must include solid image recognition capabilities. Moreover, although slightly
1365 biased towards text components, our dataset can still be considered representative of a larger app
1366 population.

1367 Furthermore, for the analysis of the precision and recall of feature matching algorithms, we only
1368 used the widgets in the main Activity of the considered apps; this set may not include widget types
1369 shown after navigation in the screens of the app. For the second experiment, we considered a single
1370 AUT (PassAndroid). Hence, the measured metrics may vary if the approach is applied to different
1371 apps.

1372

1373 To answer RQ2, we had to limit our experimentation to just two experimentation subjects for
1374 execution time reasons. To avoid cherry-picking, we considered AUTs that are commonly used in the
1375 software testing literature. To provide additional external validity to our findings, we compared our
1376 test suites with existing GUI test suites in open-source projects, leveraging a repository of GUI tests
1377 mined from GitHub. On a total of 3226 correctly identified widgets interacted in Espresso test cases,
1378 we found out that 40% of interactions were on Text-based widgets, followed by Toolbars (10.6%),
1379 different types of Buttons (8,1%), NavigationViews (4%) and ListViews (3,5%). These percentages
1380 are compatible with the test suites that we developed for our two experimental subjects.

1381 Finally, we only considered 14 emulated Android devices that are embedded in Android Studio;
1382 it is not ensured that the measured metrics can be applied to other devices, even if with equal
1383 resolution and screen size. Other context factors like, e.g., the OS version, may also have an influence
1384 on the metrics. All the considered devices also are part of the Google ecosystem and come equipped
1385 with an uncustomized Android version. Extending the experiment to other families of devices
1386 would require the use of real hardware devices or the use of third-party emulators. The first solution
1387 would require changes in the algorithms used by the VGT tools to search for the widgets in the
1388 screen of the connected device instead of the desktop environment screen; the second would require
1389 alternatives to ADB commands, to control the download of screen hierarchies and captures.

1390 *Internal validity threats.* The metrics that we measured, especially those regarding the execution
1391 time of the feature matching algorithms, strongly depend on their specific implementations. In par-
1392 ticular, it is possible that better-optimized implementations could lower execution times, especially
1393 for the feature matching techniques for which only a Python prototype was available, albeit based
1394 on well-established and widely adopted libraries such as OpenCV.

1395 We have considered only one version for the same app across multiple devices. In settings like
1396 regression testing, the actual performance may be lower than that observed due to changes in
1397 the app layout, widget, and functionalities. The internal validity of our experiments, however, is
1398 preserved since all algorithms and tools were compared on equal grounds.

1399 *Construct validity.* It is not ensured that the metrics used in this work (i.e., precision and recall for
1400 RQ1 and percentage of correct locators, interactions, and tests for RQ2) are the best possible proxies
1401 to observe the portability and the robustness to device fragmentation issues for Android VGT. We
1402 cannot ensure, for instance, that the measured metrics can correctly evaluate the fault-finding
1403 ability of test cases on different devices in a real testing scenario.

1404 *Researcher Bias* could be introduced by creating a test suite for the PassAndroid app that was
1405 made by one of the authors of the paper. However, the author was not instructed to insert specific
1406 widgets or visual locators in the test suite, neither was inclined to demonstrate a specific result.

1407 8 CONCLUSIONS AND FUTURE WORK

1409 The performance and, ultimately, the applicability of the VGT paradigm strongly depends on the
1410 robustness of the underlying image recognition algorithms. Our results show that state-of-the-
1411 art tools have limited portability across devices. A holistic approach in which the matching is
1412 simultaneously optimized for all widgets, combined with robust local feature descriptors, allowed
1413 our configurations to outperform state-of-the-art VGT tools by at least 30% in terms of correctly
1414 executed test suites.

1415 Still, much remains to be done to ensure the full portability of test suites across devices and
1416 app versions. To support future research in this domain, we release the DatAndroid dataset⁵
1417 which includes close to 100 apps rendered on multiple devices and specifically targets the issue of
1418 portability.

1420 ⁵available at <https://frankissimo.github.io/datAndroid/>

1422 As our future work, we envision to embed our matching approaches in a full-fledged VGT
 1423 tool, with capabilities of test creation, test execution/replication, and test repair in case of device
 1424 fragmentation fragility. We plan on exploiting deep learning techniques to perform a semantic
 1425 interpretation of Android GUIs screenshots, complementing and extending matching algorithms
 1426 based solely on the computation of visual similarities. Finally, we plan to validate our methodology
 1427 on different types of apps and different virtual devices, to improve the generalizability of our
 1428 results.

1430 REFERENCES

- 1431 [1] Synteda AB. 2018. EyeAutomate Documentation. [https://eyeautomate.com/wp-content/themes/EyeAutomateTheme/
 1432 resources/EyeAutomateCertifiedTesterCourse.pdf](https://eyeautomate.com/wp-content/themes/EyeAutomateTheme/resources/EyeAutomateCertifiedTesterCourse.pdf). Accessed: 2020-08-06.
- 1433 [2] Emil Alégroth. 2015. *Visual GUI Testing: Automating High-level Software Testing in Industrial Practice*. Chalmers
 1434 University of Technology, Göteborg.
- 1435 [3] Emil Alégroth and Robert Feldt. 2017. On the long-term use of visual GUI testing in industrial practice: a case study.
 1436 *Empirical Software Engineering* 22, 6 (2017), 2937–2971.
- 1437 [4] Emil Alégroth, Robert Feldt, and Pirjo Kolström. 2016. Maintenance of automated test suites in industry: An empirical
 1438 study on Visual GUI Testing. *Information and Software Technology* 73 (2016), 66–80.
- 1439 [5] Emil Alégroth, Zebao Gao, Rafael Oliveira, and Atif Memon. 2015. Conceptualization and evaluation of component-
 1440 based testing unified with visual GUI testing: an empirical study. In *2015 IEEE 8th International Conference on Software
 1441 Testing, Verification and Validation (ICST)*. IEEE, Graz, Austria, 1–10.
- 1442 [6] Emil Alegroth, Michel Nass, and Helena H Olsson. 2013. JAutomate: A tool for system-and acceptance-test automation.
 1443 In *2013 IEEE Sixth International Conference on Software Testing, Verification and Validation*. IEEE, Washington, DC, USA,
 1444 439–446.
- 1445 [7] Domenico Amalfitano, Anna Rita Fasolino, Porfirio Tramontana, Bryan Dzung Ta, and Atif M Memon. 2015. MobiGUI-
 1446 TAR: Automated model-based testing of mobile apps. *IEEE software* 32, 5 (2015), 53–59.
- 1447 [8] Domenico Amalfitano, Vincenzo Riccio, Ana CR Paiva, and Anna Rita Fasolino. 2018. Why does the orientation change
 1448 mess up my Android application? From GUI failures to code faults. *Software Testing, Verification and Reliability* 28, 1
 1449 (2018), e1654.
- 1450 [9] Marti J Anderson. 2014. Permutational multivariate analysis of variance (PERMANOVA). *Wiley statsref: statistics
 1451 reference online* 1 (2014), 1–15.
- 1452 [10] Luca Ardito, Riccardo Coppola, Maurizio Morisio, and Marco Torchiano. 2019. Espresso vs. EyeAutomate: An experiment
 1453 for the comparison of two generations of android GUI testing. In *Proceedings of the Evaluation and Assessment on
 1454 Software Engineering*. ACM, Copenhagen, Denmark, 13–22.
- 1455 [11] Luca Ardito, Riccardo Coppola, Marco Torchiano, and Emil Alégroth. 2018. Towards automated translation between
 1456 generations of GUI-based tests for mobile devices. In *Companion Proceedings for the ISSTA/ECOOP 2018 Workshops*.
 1457 ACM, Amsterdam, Netherlands, 46–53.
- 1458 [12] Ishan Banerjee, Bao Nguyen, Vahid Garousi, and Atif Memon. 2013. Graphical user interface (GUI) testing: Systematic
 1459 mapping and repository. *Information and Software Technology* 55, 10 (2013), 1679–1694.
- 1460 [13] Daniel Barath and Jiri Matas. 2017. Graph-Cut RANSAC. In *2018 IEEE/CVF Conference on Computer Vision and Pattern
 1461 Recognition*. IEEE, Salt Lake City, UT, USA, 6733–6741.
- 1462 [14] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. 2008. Speeded-Up Robust Features (SURF). *Comput.
 1463 Vis. Image Underst.* 110, 3 (June 2008), 346–359. <https://doi.org/10.1016/j.cviu.2007.09.014>
- 1464 [15] Nataniel P. Borges, Maria Gómez, and Andreas Zeller. 2018. Guiding App Testing with Mined Interaction Models.
 1465 In *Proceedings of the 5th International Conference on Mobile Software Engineering and Systems (Gothenburg, Sweden)*
 1466 (*MOBILESoft '18*). Association for Computing Machinery, New York, NY, USA, 133–143. [https://doi.org/10.1145/
 1467 3197231.3197243](https://doi.org/10.1145/3197231.3197243)
- 1468 [16] Emil Borjesson and Robert Feldt. 2012. Automated system testing using visual GUI testing tools: A comparative study
 1469 in industry. In *2012 IEEE Fifth International Conference on Software Testing, Verification and Validation*. IEEE, Montreal,
 1470 Quebec Canada, 350–359.
- 1471 [17] G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* 1 (2000), –.
- 1472 [18] Victor R Basili-Gianluigi Caldiera and H Dieter Rombach. 1994. Goal question metric paradigm. *Encyclopedia of
 1473 software engineering* 1 (1994), 528–532.
- 1474 [19] Tsung-Hsiang Chang, Tom Yeh, and Robert C Miller. 2010. GUI testing using computer vision. In *Proceedings of the
 1475 SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY,
 1476 USA, 1535–1544.

- 1471 [20] Jieshan Chen, Mulong Xie, Zhenchang Xing, Chunyang Chen, Xiwei Xu, Liming Zhu, and Guoqiang Li. 2020. Object
1472 detection for graphical user interface: old fashioned or deep learning or a combination?. In *Proceedings of the 28th ACM*
1473 *Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*.
Association for Computing Machinery, New York, NY, USA, 1202–1214.
- 1474 [21] Riccardo Coppola, Luca Ardito, and Marco Torchiano. 2019. Fragility of layout-based and visual GUI test scripts: an
1475 assessment study on a hybrid mobile application. In *Proceedings of the 10th ACM SIGSOFT International Workshop on*
1476 *Automating TEST Case Design, Selection, and Evaluation*. Association for Computing Machinery, New York, NY, USA,
1477 28–34.
- 1478 [22] Riccardo Coppola, Luca Ardito, Marco Torchiano, and Emil Alégroth. 2020. Translation from Visual to Layout-based
1479 Android Test Cases: a Proof of Concept. In *2020 IEEE International Conference on Software Testing, Verification and*
Validation Workshops (ICSTW). IEEE, Washington, DC, USA, 74–83.
- 1480 [23] Riccardo Coppola, Luca Ardito, Marco Torchiano, and Emil Alégroth. 2021. Translation from layout-based to visual
1481 android test scripts: An empirical evaluation. *Journal of Systems and Software* 171 (2021), 110845. [https://doi.org/10.](https://doi.org/10.1016/j.jss.2020.110845)
1482 [1016/j.jss.2020.110845](https://doi.org/10.1016/j.jss.2020.110845)
- 1483 [24] Riccardo Coppola, Maurizio Morisio, Marco Torchiano, and Luca Ardito. 2019. Scripted GUI testing of Android
1484 open-source apps: evolution of test code and fragility causes. *Empirical Software Engineering* 24 (2019), 1–44.
- 1485 [25] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. 2008. Image Retrieval: Ideas, Influences, and Trends of the
1486 New Age. *ACM Comput. Surv.* 40, 2, Article 5 (May 2008), 60 pages. <https://doi.org/10.1145/1348246.1348248>
- 1487 [26] Biplab Deka, Zifeng Huang, Chad Franzen, Joshua Hibschan, Daniel Afergan, Yang Li, Jeffrey Nichols, and Ranjitha
1488 Kumar. 2017. Rico: A mobile app dataset for building data-driven design applications. In *Proceedings of the 30th Annual*
ACM Symposium on User Interface Software and Technology. Association for Computing Machinery, New York, NY,
1489 USA, 845–854.
- 1490 [27] A. Developers. 2012. Ui/application exerciser monkey.
- 1491 [28] Mattia Fazzini, Eduardo Noronha de A Freitas, Shaunik Roy Choudhary, and Alessandro Orso. 2017. Barista: A
1492 technique for recording, encoding, and running platform independent android tests. In *Software Testing, Verification*
and Validation (ICST), 2017 IEEE International Conference on. IEEE, Washington, DC, USA, 149–160.
- 1493 [29] Pablo Fernández Alcantarilla. 2013. Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces. In
1494 *Proceedings of the British Machine Vision Conference*. BMVA Press, Durham, UK. <https://doi.org/10.5244/C.27.13>
- 1495 [30] K. Fukunaga and L. Hostetler. 1975. The estimation of the gradient of a density function, with applications in pattern
1496 recognition. *IEEE Transactions on Information Theory* 21, 1 (1975), 32–40.
- 1497 [31] Dan Han, Chenlei Zhang, Xiaochao Fan, Abram Hindle, Kenny Wong, and Eleni Stroulia. 2012. Understanding android
1498 fragmentation with topic analysis of vendor-specific bugs. In *Reverse Engineering (WCRE), 2012 19th Working Conference*
on. IEEE, Washington, DC, USA, 83–92.
- 1499 [32] Kristian Fjeld Hasselknippe and Jingyue Li. 2017. A novel tool for automatic GUI layout testing. In *2017 24th Asia-Pacific*
Software Engineering Conference (APSEC). IEEE, Washington, DC, USA, 695–700.
- 1500 [33] Jiajun Hu, Lili Wei, Yepang Liu, Shing-Chi Cheung, and Huaxun Huang. 2018. A tale of two cities: How webview
1501 induces bugs to android applications. In *Proceedings of the 33rd ACM/IEEE International Conference on Automated*
Software Engineering. Association for Computing Machinery, New York, NY, USA, 702–713.
- 1502 [34] Jouko Kaasila, Denzil Ferreira, Vassilis Kostakos, and Timo Ojala. 2012. Testdroid: automated remote UI testing on
1503 Android. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia*. Association for
1504 Computing Machinery, New York, NY, USA, 1–4.
- 1505 [35] Emily Kowalczyk, Myra B Cohen, and Atif M Memon. 2018. Configurations in Android testing: they matter. In
1506 *Proceedings of the 1st International Workshop on Advances in Mobile App Analysis*. Association for Computing Machinery,
1507 New York, NY, USA, 1–6.
- 1508 [36] Maurizio Leotta, Andrea Stocco, Filippo Ricca, and Paolo Tonella. 2018. Pesto: Automated migration of DOM-based
1509 Web tests towards the visual approach. *Software Testing, Verification And Reliability* 28, 4 (2018), e1665.
- 1510 [37] Ying-Dar Lin, Jose F Rojas, Edward T-H Chu, and Yuan-Cheng Lai. 2014. On the accuracy, efficiency, and reusability of
1511 automated test oracles for android devices. *IEEE Transactions on Software Engineering* 40, 10 (2014), 957–970.
- 1512 [38] Mario Linares-Vásquez, Kevin Moran, and Denys Poshyvanyk. 2017. Continuous, evolutionary and large-scale: A new
1513 perspective for automated mobile app testing. In *Software Maintenance and Evolution (ICSME), 2017 IEEE International*
Conference on. IEEE, Washington, DC, USA, 399–410.
- 1514 [39] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer*
1515 *vision* 60, 2 (2004), 91–110.
- 1516 [40] Eduardo Luna and Omar El Ariss. 2018. Edroid: A mutation tool for android apps. In *2018 6th International Conference*
in Software Engineering Research and Innovation (CONISOFT). IEEE, Washington, DC, USA, 99–108.
- 1517 [41] SS Mangiafico. 2015. An R Companion for the Handbook of Biological Statistics, Version 1.09 c, 274 p.
- 1518
- 1519

- 1520 [42] Ke Mao, Mark Harman, and Yue Jia. 2016. Sapienz: Multi-objective automated testing for Android applications. In
 1521 *Proceedings of the 25th International Symposium on Software Testing and Analysis*. Association for Computing Machinery,
 1522 New York, NY, USA, 94–105.
- 1523 [43] Matias Martinez and Bruno Gois Mateus. 2020. How and Why did developers migrate Android Applications from Java
 1524 to Kotlin? A study based on code analysis and interviews with developers.
- 1525 [44] Alessio Merlo and Gabriel Claudiu Georgiu. 2017. Riskindroid: Machine learning-based risk analysis on android. In
 1526 *IFIP International Conference on ICT Systems Security and Privacy Protection*. Springer, Cham, 538–552.
- 1527 [45] Salvador Morales-Ortega, Ponciano Jorge Escamilla-Ambrosio, Abraham Rodriguez-Mota, and Lilian D Coronado-De-
 1528 Alba. 2016. Native malware detection in smartphones with Android OS using static analysis, feature selection and
 1529 ensemble classifiers. In *2016 11th International Conference on Malicious and Unwanted Software (MALWARE)*. IEEE,
 1530 Washington, DC, USA, 1–8.
- 1531 [46] Kevin Moran, Richard Bonett, Carlos Bernal-Cárdenas, Brendan Otten, Daniel Park, and Denys Poshyvanyk. 2017.
 1532 On-device bug reporting for android applications. In *Mobile Software Engineering and Systems (MOBILESoft), 2017
 1533 IEEE/ACM 4th International Conference on*. IEEE, Washington, DC, USA, 215–216.
- 1534 [47] Maxim Mozgovoy and Evgeny Pyshkin. 2017. Using Image Recognition for Testing Hand-drawn Graphic User
 1535 Interfaces. In *Proceedings of the The Eleventh International Conference on Mobile Ubiquitous Computing, Systems, Services
 1536 and Technologies*. IARIA, Barcelona, 25–28.
- 1537 [48] Maxim Mozgovoy and Evgeny Pyshkin. 2018. Pragmatic Approach to Automated Testing of Mobile Applications
 1538 with Non-Native Graphic User Interface. In *International Journal On Advances in Software*. IARIA, Wilmington, UK,
 1539 239–246.
- 1540 [49] Maxim Mozgovoy and Evgeny Pyshkin. 2018. *Unity Application Testing Automation with Appium and Image Recognition*.
 1541 Springer International Publishing, Cham, 139–150. https://doi.org/10.1007/978-3-319-71734-0_12
- 1542 [50] Stas Negara, Naeem Esfahani, and Raymond PL Buse. 2019. Practical Android test recording with Espresso test recorder.
 1543 In *Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Practice*. IEEE,
 1544 Washington, DC, USA, 193–202.
- 1545 [51] Je-Ho Park, Young Bom Park, and Hyung Kil Ham. 2013. Fragmentation problem in Android. In *2013 International
 1546 Conference on Information Science and Applications (ICISA)*. IEEE, Washington, DC, USA, 1–2.
- 1547 [52] Alireza Sahami Shirazi, Niels Henze, Albrecht Schmidt, Robin Goldberg, Benjamin Schmidt, and Hansjörg Schmauder.
 1548 2013. Insights into layout patterns of mobile user interfaces by an automatic analysis of android apps. In *Proceedings of
 1549 the 5th ACM SIGCHI symposium on Engineering interactive computing systems*. Association for Computing Machinery,
 1550 New York, NY, USA, 275–284.
- 1551 [53] Statista. 2021. Global Google Play app downloads 2016-2019. [https://www.statista.com/statistics/734332/google-play-
 1552 app-installs-per-year/](https://www.statista.com/statistics/734332/google-play-app-installs-per-year/). Accessed: 2020-06-19.
- 1553 [54] S. A. K. Tareen and Z. Saleem. 2018. A comparative analysis of SIFT, SURF, KAZE, AKAZE, ORB, and BRISK. In *2018
 1554 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*. IEEE, Washington, DC,
 1555 USA, 1–10.
- 1556 [55] Shaharyar Ahmed Khan Tareen and Zahra Saleem. 2018. A comparative analysis of sift, surf, kaze, akaze, orb, and brisk.
 1557 In *2018 International conference on computing, mathematics and engineering technologies (iCoMET)*. IEEE, Washington,
 1558 DC, USA, 1–10.
- 1559 [56] Swapna Thorve, Chandani Sreshtha, and Na Meng. 2018. An empirical study of flaky tests in android apps. In *2018
 1560 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE, Washington, DC, USA, 534–538.
- 1561 [57] Marco Torchiano. 2020. *effsize: Efficient Effect Size Computation*. Politecnico di Torino. [https://doi.org/10.5281/zenodo.
 1562 1480624](https://doi.org/10.5281/zenodo.1480624) R package version 0.8.1.
- 1563 [58] J. Tuovonen, Mourad Oussalah, and Panos Kostakos. 2019. MAuto: Automatic Mobile Game Testing Tool Using
 1564 Image-Matching Based Approach. *The Computer Games Journal* 8 (10 2019), 215–239. [https://doi.org/10.1007/s40869-
 1565 019-00087-z](https://doi.org/10.1007/s40869-019-00087-z)
- 1566 [59] Tinne Tuytelaars and Krystian Mikolajczyk. 2008. *Local invariant feature detectors: a survey*. Now Publishers Inc,
 1567 Boston, USA.
- 1568 [60] Mikko Vesikkala. 2014-05-05. *Visual Regression Testing for Web Applications; Selainpohjaisten ohjelmistojen visuaalinen re-
 1569 gressiotestaus*. G2 Pro gradu, diplomityö; masterThesis. Aalto University. <http://urn.fi/URN:NBN:fi:aalto-201405131809>
- 1570 [61] Jiwei Yan, Hao Liu, Linjie Pan, Jun Yan, Jian Zhang, and Bin Liang. 2020. Multiple-entry testing of android applications
 1571 by constructing activity launching contexts. In *2020 IEEE/ACM 42nd International Conference on Software Engineering
 1572 (ICSE)*. IEEE, Washington, DC, USA, 457–468.
- 1573 [62] Tom Yeh, Tsung-Hsiang Chang, and Robert C Miller. 2009. Sikuli: using GUI screenshots for search and automation. In
 1574 *Proceedings of the 22nd annual ACM symposium on User interface software and technology*. Association for Computing
 1575 Machinery, New York, NY, USA, 183–192.
- 1576 [63] Hrushikesh Zadgaonkar. 2013. *Robotium Automated Testing for Android*. Packt Publishing, Birmingham, UK.

- 1569 [64] Denys Zelenchuk. 2019. Espresso and UI Automator: the Perfect Tandem. In *Android Espresso Revealed*. Springer,
1570 Cham, 165–189.
- 1571
- 1572
- 1573
- 1574
- 1575
- 1576
- 1577
- 1578
- 1579
- 1580
- 1581
- 1582
- 1583
- 1584
- 1585
- 1586
- 1587
- 1588
- 1589
- 1590
- 1591
- 1592
- 1593
- 1594
- 1595
- 1596
- 1597
- 1598
- 1599
- 1600
- 1601
- 1602
- 1603
- 1604
- 1605
- 1606
- 1607
- 1608
- 1609
- 1610
- 1611
- 1612
- 1613
- 1614
- 1615
- 1616
- 1617