A Case Study of Automating User Experience-Oriented Performance Testing on Smartphones

Gerardo Canfora, Francesco Mercaldo,
Corrado Aaron Visaggio
Dept. of Engineering - University of Sannio
Benevento, Italy
{canfora, visaggio, fmercaldo}@unisannio.it

Mauro D'Angelo, Antonio Furno,
Carminantonio Manganelli
Micron Technology, Inc.
Arzano (NA), Italy
{mdangel, afurno, amanganelli}@micron.com

Abstract – We have developed a platform named Advanced Test Environment (ATE) for supporting the design and the automatic execution of UX tests for applications running on Android smartphones. The platform collects objective metrics used to estimate the UX. In this paper, we investigate the extent that the metrics captured by ATE are able to approximate the results that are obtained from UX testing with real human users. Our findings suggest that ATE produces UX estimations that are comparable to those reported by human users. We have also compared ATE with three widespread benchmark tools that are commonly used in the industry, and the results show that ATE outperforms these tools.

***Index Terms*—user experience, mobile applications, software testing, usability, smartphone, android**

I. INTRODUCTION

Smartphones are high-end mobile phones that offer advanced computing and connectivity. Typically, today's models also serve as portable media players and cameras with high-resolution touch screens, GPS navigation, Wi-Fi, and mobile broadband access. Each phone runs a specific operating system, such as Nokia Symbian, Apple iOS, Windows Phone, and Google Android. A critical task for the producers of smartphones and mobile applications is to evaluate how the end-user perceives the responsiveness and speed of a smartphone when running applications. For this reason, user experience (UX) testing is a key phase of the production process; it helps to ensure that the end-user feels comfortable when using applications on the smartphone.

ISO 9241-210 [1] defines UX as "a person's perceptions and responses that result from the use or anticipated use of a product, system or service." Thus, UX is subjective and focuses on use. Research has shown that users respond emotionally to their computers [7]. Emotion and personality influence human behavior and interaction with the external world, including interaction with machines, computers, and software applications [2,4]. UX has subjective and objective components. The objective components include, but are not limited to speed, responsiveness, and usability; the subjective components refer to the user's moods and attitudes.

UX evaluation is usually performed using software tools (loggers, mobile agents, etc.) [2,8,13]. Automated tests consider a collection of performance metrics, including resource utilization, responsiveness, and productivity [3,5,9]. The limit of the currently proposed methods for measuring UX [4,6,11,12] is that they focus on desktop software applications or web applications; testing of applications for smartphones is targeted at the more traditional aspects concerning functionality and performance. Alternatively, UX testing can be performed using qualitative analysis [10], which consists of observing a group of users while using a product (e.g., a smartphone) and interviewing them. This approach has several drawbacks, including: high costs; strong dependency of results on the experience and attitude of users; many scenarios must be tested; amount of time needed, as it is human-based.

In order to overcome such drawbacks, Micron has developed Advanced Test Environment (ATE), a platform for automating UX tests on Android-powered smartphones. ATE supports the design, generation, and execution of test scenarios and the evaluation of test results. ATE is built around an Arduino Mega microcontroller board and a rotating platform. The Android device is placed on the rotating platform, which moves in a controlled manner, simulating the devices movements that affect the display orientation and sensors settings. An advantage of ATE is that the test designer can produce a large number of test cases in a short time because the platform generates the scenarios by combining a small number of basic patterns of actions with a script language. The expected limitation of this approach derives from ATE's use of quantitative and objective resource-monitoring measures to estimate the actual experience of the user. Objective metrics may fail to capture the actual perceptions of users influenced by subjective factors.

The paper first describes the ATE platform. It then provides an explorative case study, which led us to conclude that the ATE platform can produce results comparable to qualitative analysis with actual users.

II. ATE: MEASURING THE UX

Micron is one of the world's leading semiconductor companies. The DRAM, NAND, and NOR Flash memory produced by Micron are used in a wide array of applications, from computing networks and server applications to mobile, embedded, customer, automotive, and industrial devices, appliances, and products. The birth of the ATE platform is connected to Micron's need to automatically reproduce usage scenarios of smartphones and capture data of UX tests through a logging system to help understand how well Micron's memory products are actually performing in these devices. In this context, a test scenario is a list of actions that an Android user can perform on the device. The scenarios are designed to be as close as possible to actions that a real human user would perform on the phone. In the ATE system, the test script file can be put under configuration management and can be used multiple times. As a result, the test scenarios implemented can be saved and reused in order to allow test repeatability and automation. ATE measures how the phone is reacting to the stimuli provided. Two aspects are critical:

- the ability to design usage scenarios through a software tool that simulates the sequences of commands of the human user;
- the accuracy of the measurement since the most widespread limitations of UX testing in mobile devices is the low level of accuracy in the measurements.

The ATE solution is designed and implemented using:

- the Android Debug Bridge (ADB), which sends the phone commands that are used to recreate a user scenario;
- Android logs, which are provided by the platform to record phone feedback;
- a simple application that runs on the target phone as a daemon and implements the ATE commands necessary for the test scenario.

A. *ATE Architecture and Workflow*

Figure 1 shows the three tiers of the ATE architecture, consisting of the ATE Client, the ATE Server, and the device under test with a servo motor to handle it.

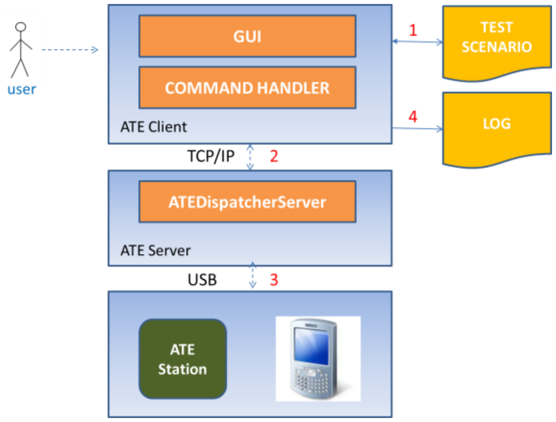

Figure 1: Three tiers of the ATE architecture

The ATE Client tier allows test scenarios to be built through a graphical user interface (GUI) (1). The test scenarios are written in the ATE native script language. This tier also includes a script engine in charge of the execution of test scenarios and a command handler to send commands to the ATE Server tier via TCP/IP. Once built, the test scenario can either be directly executed or saved in a script file. The ATE Server tier handles the commands from the ATE-Client (2). It dispatches the commands to the specific Android device or Arduino board via USB (3). To send ADB commands, the ADB client is invoked from a shell by issuing these commands, and then the ADB server manages communication between the client and the ADB daemon running on the Android device. The third tier includes the Android device tier and the Arduino board tier. On the Android device tier, the ADB daemon runs as a background process and is responsible for performing incoming commands. Some operations (e.g., killing an activity) in the ATE test scenarios are not provided by the ADB command suite. For this reason, the third tier includes an Android application designed to provide these operations. The Arduino board tier runs the ATE Station, which is a servomotor that moves the phone under test. The Arduino Servo library controls the motor. Figure 2 shows the ATE station.

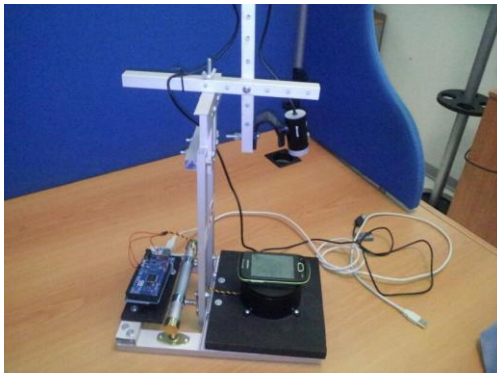

Figure 2: ATE station

Android can return information depending on the context in which the test scenario activity was launched. With that information, ATE is able to compute the execution time for starting an Android activity. When an ADB command is performed, the Android logging system publishes the related events and timestamps on the Android logs. This allows ATE to test results in a log file where the overall performance measurements for each action are stored (4).

III. Case Study Preparation

In our case study, two smartphones were used—one much slower than the other—and the reactions of the human users and the ATE were captured and recorded. Thus, we assessed whether ATE is able to perceive the same differences between the two smartphones that the human users perceived. Finally, we tested three standard benchmarks tools for measuring smartphone performance. Their evaluations were compared with the ATE's evaluations to establish which tool is able to register the users' perceptions of the UX with the highest accuracy. The case study consisted of the following steps:

Step 1: ATE was used to define, execute, and evaluate UX tests of smartphones. The devices used to perform the case study were two different configurations of the HTC™ Nexus One. The first configuration is called STOCK, which is a phone without any kernel modification. The second is called CUST-30, which is a phone with a custom kernel that allows the Android OS to reduce the device's RAM quantity by 30%.

A test scenario consists of a set of actions allowed by the specific device with a specific OS installed. We used user profiles to define the testing scenarios. In particular, we defined three profiles named Normal User, Smart User, and

Businessman. Scenarios for different types of users can have the same set of actions; the difference is in the percentage of time or occurrences in which they appear in the test scenario. For each profile, a series of services was selected, characterizing each profile's typical usage of the smartphone (e.g. voice communication, text communication, web browsing, entertainment, positioning, market).

Step 2: A demographic analysis with traditional observe-and-interview methods was performed using the same two smartphones of step 1. In order to acquire feedback from real mobile users, we created a survey consisting of six questions regarding the user's feelings and opinions about the use of a mobile device. It is in the format of a Likert questionnaire where respondents specify their level of agreement or disagreement on a symmetric agree-disagree scale for a series of statements. With the aim of assessing whether ATE metrics approximated human perception, the results of step 1 and step 2 were compared.

Step 3: Three well-known and widespread on-device benchmarks for measuring smartphone performance were applied on the two smartphones of steps 1 and 2. This helps to compare ATE with the main competing solutions in the industry. In particular, we compared ATE with Quadrant™ [16], AnTuTu Benchmark™ [15], and Smartbench™ 2011 [14]. The test results were analyzed in order to understand which tests yielded results similar to the human perceptions of the differences between the two smart-phones.

In order to perform automated testing using ATE, the devices under test were reset to factory data and attached to a host PC running the testing tool. The first step consisted of launching three different test scenarios—one for each user category—to load the device under test. The tests were launched in parallel on the three different configurations. Each test lasted two hours, for a total of six hours of testing. Once these preliminary tests were finished, three test scenarios were launched on the devices to actually obtain UX-oriented scores for each smartphone. The test scenarios were:

- cold start: an activity that isn't already loaded in the RAM is launched;
- warm start: an activity that is already loaded in the memory is launched;
- go to home: an activity is launched and then the Home activity is called, which returns the user to the Android home screen.

User participation in the experiment was completely voluntary. An e-mail was sent to Micron's R&D staff in Arzano (Naples, Italy) to ask for volunteers interested in participating in the test. The first eight persons who gave their consent to participate formed the experimental sample. Thus, the sample selection was completely random. Each device was restored to factory data before delivering it to a new participant.

The observation period lasted eight days; each day, two phones were given to two participants (one phone per person). The participants didn't know which device configuration they were going to use in order to avoid influencing them. Before starting the test, participants were required to fill out a brief survey about their behaviors using a smartphone and their habits and preferences. After that, a phone was given to them for 24 hours to let them use the device in all its aspects. After returning the device 24 hours later, the participants filled out another survey about their UX and feelings about using the device just returned.

IV. CASE STUDY RESULTS

This section discusses the data collected for each step of the case study.

Step 1: Each test is associated to a score that evaluates the UX numerically. ATE scores were calculated using the following equation:

$$ScoreATE = \mathring{a}_{i=1}^{3} \frac{1}{n(ATscript_i)} *1000 \qquad (1)$$

This equation is the summation over the inverse of the sum of the average execution times (in milliseconds) for each script used (AT_{script_i}). This solution was adopted in order to ensure that the highest value is better, just like the benchmark scores. ATE found the most emphasized differences between the two configurations: a score of 30 for CUST-37 and 42 for STOCK. Remember that the CUST-30 configuration of the HTC Nexus One has 30% less memory than the STOCK configuration.

Step 2: We analyzed the responses to the pretest survey, and then grouped the participants into three categories based on the results: Normal User (11%), Smart User (22%), and Businessman (67%). Each category was associated to a different user profile. As shown in Figure 3, the UX evaluation by different users reveals differences between the STOCK and the CUST-30 devices. A performance gap perceptible by humans, not only by automated testing tools—in particular by ATE, was found. The more significant differences between the two versions of the device were found when switching between applications and returning to the home screen (which is also a particular case of switching between applications).

Step 3: A benchmark session was performed in order to show performance differences between the devices under test. The initial state of both smartphones was the same—they were restored to factory data and reflashed with stock and custom kernels. After that, a SIM and an SD card were installed in both devices in order to run full benchmarks. The mobile network option was enabled on both phones as well. Tests were performed on several parameters, such as CPU, I/O, GPU, database, and SD card R/W features. The test performed with the benchmarks did not perceive sensitive differences between the two smartphones (figure 4): 2% for Smartbench™ and 0.1% for AnTuTu™, which is very small compared to the differences perceived by human users (12%). Quadrant™ reverses the evaluation with a score of -2%. On the other hand, ATE revealed a performance gap of about 13% indicating that the STOCK device is 13% faster than the CUST-30 device. Finally, the results of the UX evaluation performed using human participants revealed that users found a perceptible difference in terms of performance between the different configurations. The STOCK version was given a score of 4.54/5 while the CUST-30 received a score of 4.05/5. In terms of percentage, this result is very close to the result produced by the ATE evaluation.

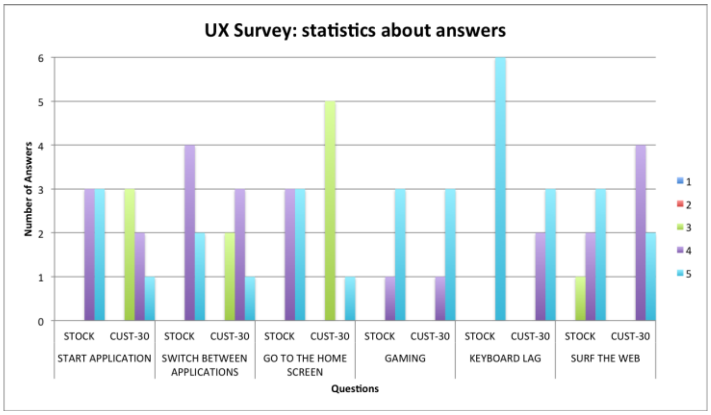

Figure 3: UX survey results

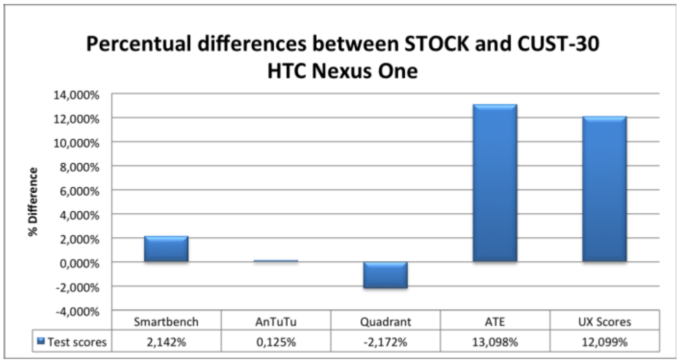

Figure 4: Performance differences among devices

V. CONCLUSIONS AND FUTURE WORK

The results of our experiment validated the user experience (UX) testing approach proposed in this paper—the feedback received from real smartphone users was in-line with the scores computed by our Advanced Test Environment (ATE) platform, highlighting the UX differences between the two test devices. While performing the experiment, another important finding was obtained: benchmark tools like Quadrant™, AnTuTu™ and Smartbench™ 2011 did not find any significant differences between the two device configurations. So, it can be asserted that benchmark tools are scarcely effective to evaluate UX for mobile devices.

The smartphone industry can benefit from a solution like ATE in a number of ways:

- the process of UX testing will become faster and cheaper when it is completely automated;
- the UX testing can be more flexible. For example, changing a scenario and running the tests again simply requires some code to be entered;
- the UX testing is perfectly reproducible; the human being is influenced by psychology, experience, and humor while the machines are not;
- there is direct traceability between the effect of poor usability with the possible causes. Because the results of tests are directly linked to the observed and measured characteristic of the phone, it is easier to find the cause of slow responsiveness.

To extend this work, it can be interesting to use our UX testing approach to test many different devices, each one with the same OS version but with different RAM amounts and memory optimization software in order to evaluate the percentage of performance loss acceptable by the users until the device becomes unusable. Another idea that is under development by the University of Sannio and the Micron validation team in Arzano is to create a tool that automatically creates a set of test scenarios compatible with some provided parameters, like quantity of CPU usage, RAM, a battery drain profile, and a user profile.

REFERENCES

[1] ISO 9241-210:2009. Ergonomics of human system interaction -Part 210: Human-centered design for interactive systems (formerly known as 13407). International Organization for Standardization (ISO), Switzerland.

[2] J. Froehlich, M.Y. Chen, S. Consolvo, B. Harrison, J. A. Landay, "My Experience: a system for in situ tracing and capturing of user feedback on mobile phones", Proc. of the 5th international conference on Mobile systems, applications and services, pp.57-70, 2007

[3] H. Gani, "The Impact of Runtime Metrics Collection on Adaptive Mobile Applications", Honours thesis, RMIT University of Melbourne, 2005

[4] S. Ickin, K. Wac, M. Fiedler, L. Janowski, J. Hong, A. Dey, "Factors Influencing Quality of Experience of Commonly Used Mobile Applications", IEEE Communication Magazine,50(4), pp. 48-56, 2012

[5] J. Kistler, M. Satyanarayanan, "Disconnected Operation in the Coda File System", ACM Transactions on Computer Systems (TOCS), 10(1), pp. 3-25, 1992

[6] W.Lai, "The study of software usability facing with user behavior model", University of South West, doctor paper, 2007.

[7] S. McDonald, R.J. Stevenson, "The effects of text structure and prior knowledge of the learner on navigation in hypertext", Human Factors, 40(1), p. 18, 1998

[8] P.Navarro, G. Pérez, D. Ruiz, "Towards Software Quality and User Satisfaction through User Interfaces", 2011 Fourth IEEE International Conference on Software Testing, Verification and Validation, pp. 415-418, 2011

[9] V. Rahimiam, J. Habibi, "Performance Evaluation of Mobile Software Systems: Challenges for a Software Engineer", Prof. Of Electrical Engineering, Computing Science and Automatic Control, 2008. CCE 2008. 5th International Conference on, pp. 346-351, 2008

[10] C. Robson, "Real World Research", Blackwell Publishing, 2011

[11] R. Rubinoff, "How To Quantify The User Experience", http://www.sitepoint.com/print/quantify-user-experience , 2004, last visit 1st August 2012

[12] A. Shye, B. Scholbrock, G. Memik, P. Dinda, "Characterizing and Modeling User Actitivity on Smartphones", Technical Report NWU-EECS-10-06, Northwestern University, 2010

[13] H. Verkasalo, "Mobile Audience Measurements in User Experience Research", Proc. Of IEEE Wireless Communications and Networking Conference (WCNC),pp. 1-6, 2010

[14] https://play.google.com/store/apps/details?id=com.smartbench.twelve&hl=it, last visit: 30th July 2012

[15] http://www.antutulabs.com/, last visit 30th July 2012

[16] https://play.google.com/store/apps/details?id=com.aurorasoftworks.quadrant.ui.standard&hl=it, last visit 30th July 2012