[POSTER] Towards Estimating Usability Ratings of Handheld Augmented Reality Using Accelerometer Data

Marc Ericson C. Santos*
Nara Institute of Science and Technology
Takafumi Taketomi†
Nara Institute of Science and Technology
Goshiro Yamamoto‡
Nara Institute of Science and Technology
Gudrun Klinker§
Technische Universität München
Christian Sandor¶
Nara Institute of Science and Technology
Hirokazu Kato‖
Nara Institute of Science and Technology

Abstract

Usability evaluations are important to the development of augmented reality systems. However, conducting large-scale longitudinal studies remains challenging because of the lack of inexpensive but appropriate methods. In response, we propose a method for implicitly estimating usability ratings based on readily available sensor logs. To demonstrate our idea, we explored the use of features of accelerometer data in estimating usability ratings in an annotation task. Results show that our implicit method corresponds with explicit usability ratings at 79% and 84%. These results should be investigated further in other use cases, with other sensor logs.

Index Terms: H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—Artificial, augmented, and virtual realities; Evaluation/methodology

1 Introduction

Designing effective augmented reality systems is challenging because there are limited established design guidelines. Often, researchers propose completely new ways of interaction between users and technology. As such, [2] explains that it is necessary to have a usability engineering approach that iteratively applies user studies to inform design. In practice, researchers ask a group of people to use a system. Often, they observe more explicit measures of usability such as errors, timing, etc. through videos, data logging and expert observers. After using the system, users give their feedback through interviews and questionnaires. In cross-sectional studies, a user would need to use systems and answer to questions multiple times. In effect, conducting user studies requires significant amount of money, time and manpower.

Lack of resources limits the scale and duration of user studies. In particular, it is challenging to conduct large-scale longitudinal studies which are necessary for some application areas. For example, in industrial work support, many workers would use the system for weeks before we can observe improvements in collaboration and productivity. In learning support, multiple students need to use the system simultaneously during class hours. Moreover, learning needs to be observed over the duration of month-long courses.

In response, we explore whether in some situations, more implicit user studies can be conducted – especially when handheld devices are involved. As an example, we estimate the usability of one specific function in augmented reality – text annotation, using one specific sensor – the accelerometer.

*e-mail: chavez-s@is.naist.jp
†e-mail:takafumi-t@is.naist.jp
‡e-mail:goshiro@is.naist.jp
§e-mail:klinker@in.tum.de
¶e-mail:sandor@is.naist.jp
‖e-mail:kato@is.naist.jp

2 Proposed Method

Unique usability issues arise when handheld devices such as smartphones and tablets are used for augmented reality [8]. One issue that significantly influences the usability of handheld augmented reality or HAR is its manipulability – the ease of handling the HAR system [7]. HAR requires the user to grip, move and pose the device in unconventional ways. As such, the device's accelerometer data might contain implicit information on usability [6].

HAR systems for work support, learning support and other application areas may include common functions such as annotating text, object positioning, etc. To estimate usability for each function, we propose the use of automatically generated sensor data such as accelerometer data, gyroscope data, etc. that are logged while a user uses each function. To analyze this information, we recommend labelling a small part of this data set with manually gathered usability ratings thereby forming a gold standard. The usability rating from the rest of the data set can then be estimated by comparing with the gold standard. Using this method, we can minimize the number of times users need to answer questionnaires explicitly.

Finally, based on the user study, researchers could improve their system or provide some additional training on using the system, especially for users having difficulty.

3 Experiment

We demonstrate our proposed method for the text annotation function of a handheld augmented reality system. This function introduces unconventional gestures with a tablet. In this scenario, users create a SLAM map by swinging the device from side-to-side. They then create several 3D-registered labels by pointing the device to the target object, tapping on the screen, and typing the text.

3.1 Platform

We implemented a HAR authoring tool for text annotations on real objects as shown in Figure 1. It runs on iPad 2 tablets and it uses SLAM point clouds for tracking. To create the point cloud map, the user needs to move the device from side-to-side. After the system detects enough points, the user can add text annotations on to the scene. While the system is in use, it records accelerations in X, Y and Z directions in the background at 60 logs per second.

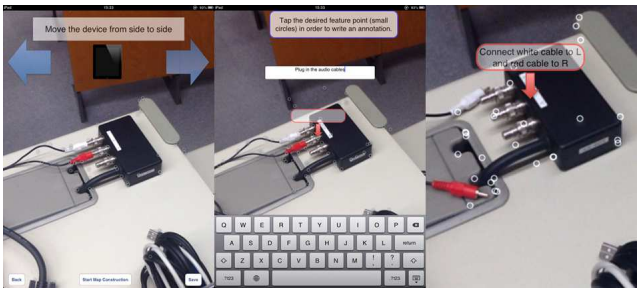

Figure 1: Screenshots of AR System for Annotating Text

3.2 Instruments

We use the System Usability Scale (SUS), a valid and reliable usability questionnaire. It aggregates usability ratings into a single score [3]. Based on this score, we can judge if a system is good enough for the target user to accomplish a specific task. In our work, usability refers to how well target users can use a functionality of a system to accomplish a specific task [4].

3.3 Procedures

We recruited 23 voluntary participants (22 to 42 years, 15 male, 8 female). All of them use handheld devices daily. Fourteen have experienced using an augmented reality system at least once. We demonstrated to them how to use the system. We then asked them to add English translations to a Japanese rice cooker and trivia on a Philippine paper bill, as shown in Figure 2. We did not set a time limit and the participants could opt out during the experiment. After the task, we asked the participants to answer the SUS.

Figure 2: Scenario and Task

3.4 Data Analysis

For each participant's accelerometer log (X, Y, Z), we extracted time and frequency domain feature sets recommended in [5]. Table 1 lists the description of the features and the total number of features in the set. We then labeled each feature set as either having "good" or "bad" usability based on the interpretation of the SUS in [3]. For each labeled set of features, we trained J48 decision trees and random trees using WEKA 3.6.[1] We chose decision tree classifiers because they capture our intuition: If the user moves the device too quickly or too slowly, or if they frequently move the device in the wrong direction, they will perceive difficulty. Otherwise, they will find the system easy to use. Finally, we evaluate the tree models using leave-one-out cross validation [1].

We used leave-one-out cross validation because we only had 23 participants. Our proposed method is for a big sample size N. For example, we can create a tree model from the first 22 participants, then use the model to estimate the usability ratings of the 23rd to the Nth participant. Using leave-one-out means that we take 22 participants to build the model, then estimate the usability rating of the remaining person. This is done iteratively wherein all 23 participants assume the role of the remaining person.

Table 1: Summary of Time and Frequency Domain Features

Features	Description	N
Mean and SD	Mean and standard deviation	6
Multiple Statistics	Mean, standard deviation, median, 25th and 75th percentile	15
Spectral Energy	Sum of squared FFT coefficients	3
FFT Magnitude	Magnitude of first five components of FFT analysis	15
Combination	Multiple statistics, spectral energy and FFT magnitude	33

[1] http://www.cs.waikato.ac.nz/ml/weka/

4 Results and Discussion

Four of the 23 accelerometer logs were missing due to logging malfunction. As such, we only have 19 samples. Seven rated the system "good," whereas 12 rated "bad." As shown in Table 2, the first four feature sets were either moderately lower or higher than 50%. However, combining these time and frequency domain features boosted the accuracy of classification by around 20%. Based only on accelerometer data, we can estimate the usability ratings of users in a simple authoring scenario at a rate of 79–84%. To improve the accuracy, other features of accelerometer data can be explored. Moreover, using device orientation logs and navigation logs may also contribute to better accuracy.

Table 2: Correctly Classified Instances (%)

Features	J4.8 Tree	Random Tree
Mean and SD	53%	58%
Multiple Statistics	42%	68%
Spectral Energy	47%	58%
FFT Magnitude	58%	63%
Combination	**79%**	**84%**

5 Conclusions

We demonstrated our proposed method for estimating usability ratings in one specific use case with one specific sensor that is readily available in smartphones and tablets. With this approach, we can inexpensively conduct large-scale longitudinal studies with handheld augmented reality systems. Such studies are necessary for generating insights on how we can improve augmented reality systems and leverage on them more effectively. It is important to further investigate our results in other use cases with other sensor logs. In addition, estimating more challenging classifications offered by the SUS such as letter grades A to F must be investigated.

Acknowledgements

This work was supported by the Grant-in-Aid for JSPS Fellows, Grant Number 15J10186.

References

[1] A. Dalton and G. OLaighin. Comparing supervised learning techniques on the task of physical activity recognition. *IEEE Journal of Biomedical and Health Informatics*, 17(1):46–52, 2013.

[2] J. L. Gabbard and J. E. Swan. Usability engineering for augmented reality: Employing user-based studies to inform design. *IEEE Transactions on Visualization and Computer Graphics*, 14(3):513–525, 2008.

[3] J. R. Lewis and J. Sauro. The factor structure of the system usability scale. In *Human Centered Design*, pages 94–103. Springer, 2009.

[4] J. Nielsen. *Usability Engineering*. Elsevier, 1994.

[5] S. J. Preece, J. Y. Goulermas, L. P. Kenney, and D. Howard. A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data. *IEEE Transactions on Biomedical Engineering*, 56(3):871–879, 2009.

[6] A. Sahami Shirazi, N. Henze, T. Dingler, K. Kunze, and A. Schmidt. Upright or sideways?: Analysis of smartphone postures in the wild. In *Proceedings of the International Conference on Human-computer Interaction with Mobile Devices and Services*, pages 362–371, 2013.

[7] M. E. C. Santos, J. Polvi, T. Taketomi, G. Yamamoto, C. Sandor, and H. Kato. A usability scale for handheld augmented reality. In *Proceedings of the Symposium on Virtual Reality Software and Technology*, pages 167–176, 2014.

[8] E. Veas and E. Kruijff. Vesp'r: Design and evaluation of a handheld ar device. In *Proceedings of the International Symposium on Mixed and Augmented Reality*, pages 43–52, 2008.