An Usability Study of Continuous Biometrics Authentication

Geraldine Kwang[1], Roland H.C. Yap[1], Terence Sim[1], and Rajiv Ramnath[2]

[1] School of Computing, National University of Singapore, Singapore
u0407321@alumni.nus.edu.sg, {ryap,tsim}@comp.nus.edu.sg
[2] Temasek Laboratories, National University of Singapore, Singapore
tslrr@nus.edu.sg

Abstract. We present an usability study for a bi-modality Continuous Biometrics Authentication System (CBAS) that runs on the Windows platform. Our CBAS combines fingerprint and facial biometrics to authenticate users. As authentication is continuous, CBAS constantly contributes a computational overhead of up to 42% to the computer system. This usability study seeks to investigate (a) whether this overhead will have an impact on the performance of users to complete tasks; and (b) whether the users deem the responsiveness of the system to be acceptable. The results of our study are encouraging, indicating that the runtime cost of a CBAS system has no measurable statistical impact on the task completion by users. We found that user acceptance of CBAS to be good and they did not perceive the CBAS to degrade system response. This suggests that continuous biometrics for authentication is viable – the CBAS benefits outweighs system impact drawbacks.

1 Introduction

Most computer systems require authentication in order to use the system. Traditionally, the authentication is done at sign-on or login time either with a password and/or biometric authentication. Combining a password with biometrics can give additional assurance that the user logging in is the authorized user rather than an intruder. However, once the user has successfully signed-on or logged-in, most systems assume that the user continues to be legitimate.

In critical or high assurance environments, we would want to ensure that the user continues to be the legitimate user. Thus, authentication of the user needs to performed in a "*continuous*" fashion over the entire time interval when the user is actively using the system. It is also desirable that the authentication process be transparent so that users do not need to do anything special, e.g. periodic password challenges are not transparent. Continuous authentication using biometrics fits these needs since many biometrics can be monitored in a non-intrusive way at a sufficient frequency to approximate continuous authentication.

Ideally, a Continuous Biometric Authentication System (CBAS) should be transparent so that the user is hardly aware of its presence. This paper investigates the usability issues for a Windows-based CBAS. We believe this is the

M. Tistarelli and M.S. Nixon (Eds.): ICB 2009, LNCS 5558, pp. 828–837, 2009.

first such study to evaluate the viability of CBAS on a general user population. Although the usage of a CBAS can be transparent, the overheads of continuous biometric verification and fusion can be significant which impacts on usability. We study the user perception of CBAS on first time users and whether the runtime overhead of CBAS affects their performance on common Windows tasks.

Our study on a moderate sized user population shows that a CBAS is quite acceptable to users. For interactive programs, such as Microsoft Office productivity applications, the system overhead of the CBAS, while significant, did not cause significant statistical difference on user performance. Although not the focus of this paper, we also show that the false reject rates for continuous biometrics are quite low. This paper gives the first successful use of continuous biometrics with a realistic user study which shows that continuous biometrics is usable for a general user population and also the viability of continuous authentication as security mechanism. Our results shows the potential of CBAS to achieve transparent continuous authentication without causing significant impact on overall usability.

2 Related Work

Recently there has been interest in continuous authentication. In response to the 11 September 2001 attacks, two design proposals using continuous biometrics authentication to safeguard the aircraft cockpit against unauthorized control were proposed by Carrilo [1].

Altinok and Turk [2] experimented with using voice, face, and fingerprint biometrics for continuous authentication. They articulated two key issues in continuous authentication, that is, the need for integration across both modality and time, and the requirement that the system must be able to determine "authentication certainty" at any point in time, even in the absence of observations. However, their work focused only on multimodal fusion and they did not study a real system with realistic workloads – their experiments used only simulated individuals because of the difficulty of getting real users.

In our previous work [3], we developed a bi-modal continuous biometrics authentication system integrated with the Linux kernel. This paper is based on a newer version of the CBAS which has been integrated with Windows.

Existing biometrics usability studies are based on one-off authentication systems which focus on the system's ability to enroll and verify individuals effectively. One-off systems are not relevant in a continuous authentication setting. Furthermore, CBAS has the additional factor of continuous runtime overhead which can impact on usability. We believe that this is the first paper to measure CBAS usability in realistic environment and task setting using task completion time [4] as an objective performance metric.

3 The Windows CBAS

This paper does not go into the details of the Windows XP-based CBAS system. Instead, we refer the interested reader to our previous paper [3] in which the

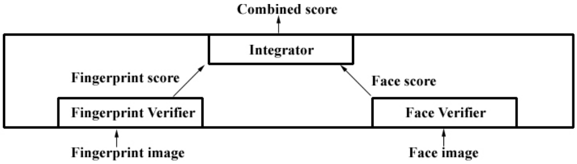

Fig. 1. CBAS architecture

decision-making method (using a Hidden Markov Model (HMM)) is described for the Linux operating system. We largely implemented the same method in Windows XP, with appropriate changes to take into account the differences in the operating system internals [5]. Here, we briefly describe the overall CBAS architecture, which suffices to understand the rest of the paper.

Our CBAS uses a normal webcam and a special mouse with a fingerprint sensor located at where the thumb would normally be placed. Face images captured by the webcam (at one image per second) are sent to a *Face Verifier* to compute a face score, while thumbprint images captured asynchronously by the fingerprint mouse (at one image every two seconds) are sent to the *Fingerprint Verifier* to compute a fingerprint score. Both scores are fused by the *Integrator* to produce a *Combined Score*. Fig. 1 shows these three components. The *Combined Score*, which reflects the probability of the continued presence of the user, determines whether the user is legitimate or an imposter. When CBAS determines the user is an intruder, it can be configured to disable the keyboard and mouse (thus locking out the user) to prevent further usage of the computer.

As our CBAS system fuses video and fingerprint biometrics, it allows for transparent monitoring. However, the challenge is that the continuous authentication requires frequent verification of both biometrics (between 0.5-2s) which itself leads to substantial overheads. We ran micro- and macro-benchmarks and found that overheads ranged from 26-42% which is consistent with [3]. This overhead is significant and highlights the need for usability studies on the performance impact of CBAS.

4 CBAS Experimental Design

The focus of this paper is to determine whether the overheads of CBAS will have an impact on overall usability on a user population. Participants were invited to our usability lab to carry out four work tasks for 1.5 to 2 hours on a workstation that had CBAS running in the background. They were told that the purpose of their participation was to help us try out a novel continuous biometrics authentication system.

Participants are first enrolled into the CBAS to train the face and fingerprint verifiers. After enrollment, participants were instructed to open the CBAS application called *BioMonitor* before beginning their tasks. They were informed that continuous authentication starts the minute they opened the BioMonitor application. In reality, the BioMonitor could be either the actual CBAS BioMonitor or a placebo BioMonitor which resembled and behaved exactly like the real one. The overhead of the placebo BioMonitor was negligible. This allows us to subject the same participant to two different test conditions: (i) with the real BioMonitor running; and (ii) with the placebo BioMonitor running.

Out of the four tasks, the last task actually repeated a task similar to the first. Participants were asked to leave for a five-minute break after the third task was finished. While they were outside the lab, the version of the BioMonitor used was switched with the other BioMonitor (real with placebo, and vice versa). The experiment is designed to measure the effect of the real BioMonitor on task completion time: (i) for the real BioMonitor group, the effect of removing the BioMonitor overhead; and (ii) for the placebo BioMonitor group, the effect of adding the BioMonitor overhead. In the experimental setup, the placebo and real BioMonitor look alike, thus any difference in task completion times between the real and placebo can be attributed to the BioMonitor overhead [6].

Only one of the three tasks was repeated because it was difficult to make participants repeat more than one task; they may get bored, tired, annoyed. The first task was chosen to be the repeat task since it was done furthest away in time. According to Chambliss [7], repeating a task raises the problem of learning effects transfer. This means that the participant will become more familiar with the task after each repetition and therefore would be able to finish it more quickly. However, the repeated task should be one which should take about the same time, hence, small changes were made to the contents of the task (transposed text, different values and pictures in the sample task documents) to reduce learning effects. Despite this, we found that we still had to adjust for some learning effects (see Sec. 5.1).

4.1 Task Set

Tasks were designed to resemble possible real-life office work using Microsoft Word, Excel and PowerPoint. They were predefined so that users were only required to replicate the sample documents which they were given. These tasks used applications which were familiar to the participants. To accommodate participants with varying levels of expertise, the tasks required only the use of basic application features.

Each task was given in the form of a realistic scenario where the participant played the role of an employee having to do some work for his supervisor. Not only does this help the participant take on the role of actual end-users, it also means that the experiment results will be a better predictor of CBAS's performance in the workplace [8]. Participants may also find it easier to "stay in role" and overcome any latent hesitation and self-consciousness if the scenario reflected familiar situations, with realistic reasons (motivation) for performing the tasks [8]. During the pre-experiment briefing, participants were also reminded to perform the tasks as they normally would.

According to Dumas and Redish [4], usability testing is an emotional experience for participants. Many of them will be tense and they will need frequent breaks to retain their concentration if they have to perform for more than a few hours. Hence, participants were asked to play computer games or watch short videos for about 10 minutes in between tasks.

We permuted the three tasks (Word, PowerPoint, Excel) to obtain six possible task order permutations. Each participant was then randomly assigned one permutation. The different permutations allow us to determine whether or not the results will vary when different types of tasks are repeated. It also reduces the chance that task order will affect completion time.

To keep track of the time-taken to complete individual tasks accurately, we used *Process Logger* [9] to track when programs on Windows are started and finished. It was chosen because of its low overhead (less than 1%). The timings used were real time since we want to measure task performance and overhead rather than CPU usage.

Dumas and Redish [4] highlight two problems related to timing how long it takes to complete tasks: (i) the participant has completed the task but is not ready to say that the task is done; and (ii) the participant says that the task is done, but in reality it is not. As such, the Experiment Administrator has to play the role of a moderator (to consistently apply the policy on when a task is considered done) and was present at all times to:

1. Observe the participants and intervene if they appeared to be having trouble with an application feature (e.g. inserting a table into a slide).
2. Ensure that the tasks were completed correctly (accuracy) before allowing them to continue. Since we want to time the tasks separately, we need a way to get participants to stop between tasks [4]. Hence, participants were given one task at a time. After checking that the task had been completed satisfactorily, the Experiment Administrator would then instruct the participant to quit the current application before beginning the next task.
3. Stop them if they have finished the task but still want to check their work. This is because prolonged checking should not be considered as task time.

4.2 Participants

The experiment was open to all students from our university since Microsoft Office 2003 is the most commonly-used productivity suite for their regular academic work. Over a two-week period, we had a total of 58 volunteers participating in the study coming from various faculties (29 males and 29 females).

5 Experimental Results

All participants, regardless of whether they used the real or placebo BioMonitor, underwent enrollment to train CBAS to recognize their face and fingerprint biometrics. This was needed for the placebo participants since their interaction with the system had to be identical to that of the real BioMonitor participants. Besides, they were also tested with the real BioMonitor for the final task.[1] The

[1] Recall that the placebo and real BioMonitor participants switch roles for the repeat (fourth) task.

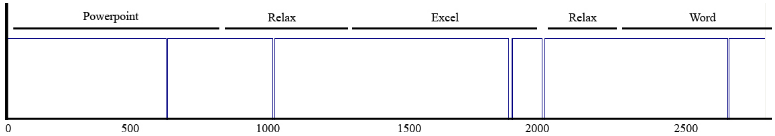

Fig. 2. A plot of *Combined Score* (vertical axis) vs. time (horizontal axis, in seconds) for a user. The dips which drop to 0 in the plot are False Rejects, i.e. where the *Combined Score* drops below the legitimate user/imposter threshold. The horizontal lines at the top indicate which application the user is running at the given times.

mean time for face enrollment was 101s, and that for fingerprint enrollment was 43s. Only 24% needed to repeat face enrollment, and only one person had to repeat fingerprint enrollment. Since all participants were first time users of the system, the enrollment process was rather smooth with the mean total enrollment time of 144s.

Although this study is primarily concerned with task performance under a CBAS, we also report the False Reject Rates (FRR). The FRR is the percentage of time the system falsely determines that the true user is an imposter. The BioMonitor continuously identified 48% of the participants correctly all the time (FRR of 0%) and the remainder (52%) had a non-zero but small FRR. For the real BioMonitor group, the mean FRR was 0.86%. In the case of the placebo BioMonitor group, the mean FRR was a little larger at 3.1%. This is not unexpected since the placebo group use the real BioMonitor during the fourth task which was only for a short period. Overall, the FRR is low, which suggests that most of the time the CBAS is *correctly identifying the user.*

In many cases, there were periods between the switching of tasks where no biometric observations were obtained and thus the BioMonitor could not verify the presence of the user. In other cases, users needed some time to settle down. Fig. 2 is an example of a graph of a user session where FRR > 0. There are a few small blips on the graph which is reasonable over the total period of 2800s. We expect that that the user might not always be using the mouse or have his face adequately captured by the webcam.

Two remarks about FRR and FAR (False Accept Rates) are in order: (1) Because we are primarily concerned with the impact of overheads on task completion time, we do not lock the user out when the *Combined Score* falls below the threshold. Doing so necessitates re-authentication, which will introduce delays not directly caused by overheads. Moreover, if the FRR is high, frequent lock-outs will frustrate the user, and may even cause him to abort work altogether. While these effects are worth studying, it is not the primary goal of this paper. (2) It is well-known that the FRR can be traded off with the FAR. Again, while it may be interesting to study such trade-offs, this is not the goal of our current work. Thus no FAR is reported here.

5.1 Accounting for Learning Effects

We first studied whether there are learning effects (see Sec. 4) in the completion time for repeating the last task. We invited 28 more participants for this learning

Table 1. Paired t-test results for learning effects

Task	N	Paired Differences					t	df	Sig- 2 tailed
		Mean Time	Std.Dev	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Higher			
Word	10	59.100	48.884	15.459	24.130	94.070	3.823	9	.004
Powerpoint	8	233.500	140.495	49.672	116.043	350.957	4.701	7	.002
Excel	10	212.100	132.752	41.980	117.135	307.065	5.052	9	.001

effects experiment. These participants did not have to go through the enrollment process and they used a normal Windows setup with neither real nor placebo BioMonitor running.

Paired t-tests were carried out on the difference between the first versus the second completion time for all task types. Table 1 shows a summary of the paired t-test results. The mean times are measured in seconds. Our paired-samples t-test analysis indicates that at the $p < .01$ level:

1. For the 10 participants who repeated the *Word Task*, the mean time taken to complete it the first time ($518.20s$) was *significantly higher* ($p = .004, t(9) = 3.823$) than the mean time to do it the second time ($459.10s$).
2. For the 8 participants who repeated the *PowerPoint Task*, the mean first task completion time ($775.38s$) was *significantly higher* ($p = .002, t(7) = 4.701$) than the mean second task completion time ($541.88s$).
3. For the 10 participants who repeated the *Excel Task*, the mean first task completion time ($917.70s$) was *significantly higher* ($p = .001, t(9) = 5.052$) than the mean second task completion time ($705.60s$).

This means that despite our best efforts to reduce learning effects (the repeat task is similar to the first so that the task completion time should be similar), participants were still able to repeat the tasks more quickly for all task types.

To account for learning effects, the mean time differences (Word = 59.1s, PowerPoint = 233.5s, Excel = 212.1s) were used to adjust the completion time of the repeat tasks for the placebo and real BioMonitor groups. Although this is admittedly crude, it is unclear how else to account for the learning effect so that the participants do not suffer from experimental fatigue and other complications (more repetitions would have to account for more learning effects).

5.2 Completion Time with and without CBAS

To determine whether the BioMonitor had any effect on completion time, we carried out paired t-tests on the difference between the completion time of the BioMonitor group versus the placebo group ($n = 58$). The combined paired t-test results in Table 2 show that at the $p < .01$ level:

- There was *no significant difference* between time taken to complete the *Word* task with and without the real BioMonitor ($p = .135$)[$t(21) = -1.554$, $p > .05$]).

Table 2. Paired t-test results for task completion time

Task	N	Paired Differences: BioMonitor - Placebo Completion Time					t	df	Sig- 2 tailed
		Mean Time	Std.Dev	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Higher			
Word	22	-21.955	66.278	14.131	-51.341	7.432	-1.554	21	.135
Powerpoint	18	121.500	320.883	75.633	-38.071	281.071	1.606	17	.127
Excel	18	17.056	171.686	40.467	-68.322	102.433	.421	17	.679

- There was *no significant difference* between time taken to complete the *PowerPoint* task with and without the real BioMonitor ($p = .127)[t(17) = 1.606, p > .05]$).
- There was *no significant difference* between time taken to complete the *Excel* Task with and without the real BioMonitor ($p = .679)[t(18) = .421, p > .05]$).

We see that there is no significant difference in completion times for all tasks. This gives evidence that even though the CBAS has significant impact on system overhead, for the interactive tasks tested in this experiment, we cannot conclude that there is any significant impact on task performance.

6 Post Usability Experiment Survey

As part of our usability study, we were also interested in finding out the subjective satisfaction of the participants with using CBAS. This is because there may be psychological factors at play, even with small system overheads the users might feel that the system is less usable or less responsive. On the other hand, there might be significant system overheads, yet the users could be quite happy with using the system.

Since human perception is to a large extent subjective, post test surveys are an essential part of usability studies because they collect subjective evaluation data about the usability of the system [8]. Our survey was self formulated; using five-point Likert scales for ratings and open-ended questions for participants to share their views in their own words. Due to space constraints, we list only four sample survey questions here, and discuss the findings in the next section:

- I felt comfortable using this system. (rating)
- Overall, I am satisfied with the responsiveness of the system whenever I type using the keyboard. (rating)
- Overall, I was comfortable using the fingerprint mouse throughout the session. (rating)
- Please share with us how you feel this biometrics monitoring security system can be improved. (open-ended)

7 Discussion

Our experimental study shows that there are no significant differences in the time taken to complete a task by the same participant, whether or not the CBAS is continuously running in the background. We were pleasantly surprised by these results as we had expected that the high CBAS overhead (up to 42%) would have more impact in increasing the overall task time.

Analysis of the answers from the post experiment survey also show that participants were generally satisfied with the system's responsiveness to mouse and keyboard input. We had similar positive responses to questions on comfort level, satisfaction with the CBAS and overall ease of use. Thus, both the experimental results and the responses from the post experiment survey are consistent. We conclude that the CPU overhead of running the CBAS did not have a significant effect on the users' task performance or in their perception of system usability for all the tasks tested in the experiment.

We wish to point out that we do not claim that there is no difference between having the CBAS running versus not having it. System measurements show there is system impact but the question is whether this is significant given a realistic system use scenario. As CBAS is for humans, we focus on appropriate human-centric measures. Rather than CPU-bound tasks, we focus on interactive tasks. Furthermore, we measure our results based on the user performance metrics rather than system overhead.

When asked to share with us what they did not like about CBAS, many participants said that because the webcam was "too obvious", it made them feel uncomfortable. They also suggested that it would be better to have the camera hidden away. This shows that the discomfort may be not really be due to the fact that they were subjected to surveillance, but more so due to them being able to see webcam. This suggests that CBAS systems need to have more discreet camera design/placement. As for the fingerprint mouse, all participants said they were comfortable using it.

Finally, although our objective is to determine whether or not CBAS has an impact from a user's perspective, we also have some interesting results on enrollment and FRR. Even though our participants were all first time users to the CBAS, we had no problems with enrollment and the mean time needed was small, just 144s. We did expect the FRR to be non zero. However, in many cases, it turned out to be either zero for the entire session or for most of the session. Both of these findings also suggest that actual implementation and deployment of a CBAS may be easy.

8 Conclusion and Future Work

Although there have been proposals for CBAS design and prototype systems, the important human factor issues have not been previously addressed. We show using a moderately large scale usability experiment on real users that the effect of CBAS on users is small. They can perform tasks just as well even with a large CBAS system overhead. From a user perspective, the presence of the CBAS is not readily perceived. It is also relatively easy to enroll new users to the CBAS. The authentication error from continuous authentication is also small as the false reject rate is under 1%. Overall our results show that the key objective of a transparent CBAS is realizable. Furthermore, we did not experience any user acceptance problems.

In the near future, we plan to: (a) gather participants from a wider age group / education background, so as to better generalize our findings to the larger population; (b) allow for multitasking to better approximate the working environment in the real world (our current study restricts users to one task at a time); (c) account for the novelty factor of using biometrics, which could have caused users to be more tolerant of any degradation in system performance; and (d) explore whether multi-core CPUs can reduce system overheads.

Acknowledgments

We acknowledge the support of Microsoft Research Asia (R-252-000-260-720) and the Defence Science Technology Agency of Singapore (R-252-000-261-422).

References

1. Carrillo, C.: Continuous biometric authentication for authorized aircraft personnel: A proposed design. Master's thesis, Naval Postgraduate School (2003)
2. Snelick, R., Indovina, M., Yen, J., Mink, A.: Multimodal biometrics: issues in design and testing. In: International conference on Multimodal interfaces, pp. 68–72 (2003)
3. Kumar, S., Sim, T., Janakiraman, R., Zhang, S.: Using continuous biometric verification to protect interactive login sessions. In: Annual Computer Security Applications Conference, pp. 441–450 (2005)
4. Dumas, J., Redish, J.: A Practical Guide to Usability Testing. Greenwood Publishing Group Inc., USA (1993)
5. Yap, R., Sim, T., Kwang, G., Ramnath, R.: Physical access protection using continuous authentication. In: IEEE Conference on Technologies for Homeland Security, pp. 510–512 (2008)
6. Alreck, P., Settle, R.: The Survey Research Handbook, 3rd edn. McGrawHill Publishers, New York (2004)
7. Chambliss, D.: Making Sense of the Social World. Sage Publications Inc., USA (2003)
8. Rubin, J.: Handbook of Usability Testing: How to Plan, Design and Conduct Effective Tests. John Wiley and Sons Inc., USA (1994)
9. Process logger, http://keleos.h11.ru/proclog