Towards understanding the effects of practice on behavioural biometric recognition performance

E. Haasnoot[1] J.S. Barnhoorn[2] L.J Spreeuwers[1]
R.N.J Veldhuis[1] W.B. Verwey[2]
University of Twente
[1]EWI, Data Science Group [2]BMS, Group CPE
Drienerlolaan 5, 7522 NB Enschede
e.haasnoot@utwente.nl

***Abstract*—Behavioural biometrics looks at discriminative features of a person's measurable behaviour, which is known to show high variance over long stretches of time. In psychology, a significant portion of this behavioural variance is explained by an individual improving their skill at performing behaviours, mostly through practice. Understanding what the effects of practice are on biometric recognition performance should allow us to account for much of this variance, as well as make individual behavioural biometric studies easier to compare [15]. We hypothesize that more accumulated practice will lead to both more stable and increased recognition performance. We argue that these are significant effects and show that practice in general is under-investigated. We introduce a novel method of analysis, the Start-to-Train Interval (STI)/Train-to-Test Interval (TTI) contour plot, which allows for systematic investigation of how recognition performance develops under increased practice. We applied this method to three data sets of a Discrete Sequence Production (DSP) task, a task that consists of repeatedly (500+ times) typing in a simple password, and found that more practice both significantly increases recognition performance and makes it more stable. These findings call for further investigation into the effects of practice on recognition performance for more standard behavioural biometric paradigms.**

I. Introduction

Behavioural Biometrics (BB) in the context of security and authentication looks at measured behaviour in order to make decisions about a person's identity. Many different types of behaviour can be used, e.g. flutists show enough idiosyncracies in their play to allow recognition by the movement of their flute in space [3] and amateur radio operators can learn to recognize other operators by their telegraphic style, or "fist", even before direct identification is received [12]. Both are examples of idiosyncratic behaviour by experts, similar in level of skill to the professional typists [27] which the earliest behavioural biometric studies focused on [12], [24].

Contemporary behavioural biometric studies focus both more on broader populations, which will include people with skill levels ranging from novice, to intermediate to expert, and on developing new task paradigms (e.g. [8], [20]), where initially people of expert skill might not even exist. In both cases, we should expect significant skill development to happen in the lifetime of a behavioural biometric application, with a major driver being (deliberate) practice [11]. Practice should have significant effects on an individual's behaviour, but the effects of practice on behavioural biometric recognition performance are not yet well understood.

A. Practice & Motor Skill Learning

We look to (cognitive) psychology literature, and specifically motor skill learning literature, as a first step towards understanding such effects. In motor skill learning, skill improvements are defined as wholesale changes in the location and shape of a Speed-Accuracy Trade-off Function (SATF). SATFs denote a fundamental trade-off; increased speed of execution will mean loss of accuracy, and vice versa [13]. From this definition of skill improvement, two possible effects of practice on recognition performance become apparent. One, an increasing convexity of the SATF under improving skill will mean fewer SATF-related behavioural adaptions, resulting in lower within-subject behavioural variance for higher skilled individuals, which should mean improved recognition performance down the line, see Figure 1a. Two, the shifts in location of the SATF slow down as skill improves, indicating that the range of behaviours related to the SATF changes slower, which should result in more stable recognition performance with more accumulated practice, see Figure 1b.

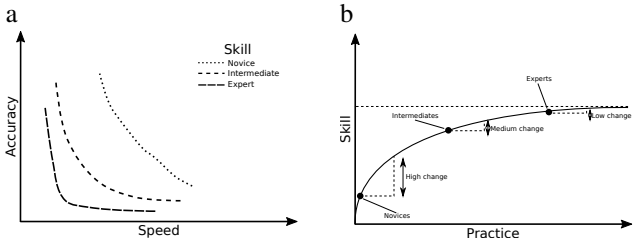

Table I
(A) A STYLIZED EXAMPLE TO SHOW WHAT SATFS MIGHT LOOK LIKE FOR INDIVIDUAL OF NOVICE, INTERMEDIATE AND EXPERT SKILL, BASED ON FINDINGS BY [18]. NOTE THE CHANGING SHAPE AND LOCATION OF THE SATF. (B) AN EXAMPLE OF A POWER LAW OF PRACTICE. NOTE THE RATE OF SKILL CHANGE BECOMING LOWER FOR THOSE WITH HIGHER INITIAL SKILL/MORE ACCUMULATED PRACTICE.

B. Practice & Behavioural Biometrics

If the above hypothesized effects are true, behavioural biometric samples are fundamentally ordered. Collecting a single sample always involves some practice and thus can improve skill, which will change the behavioural characteristics underlying the samples and thus affect the next instance of sample collection. How much will depend both on initial skill levels, as well as on how much improvement in the task is feasible. E.g., a new behavioural biometric paradigm will show very low initial skill levels, which results in high order dependency (i.e., there is much to improve skill-wise), whereas sample order for behavioural biometric modalities such as gait is less important, as the general population can be expected to be skilled walkers. Due to high likely sample order dependence for new behavioural biometric applications, we should thus present our results across some dimension of practice. We assume repetition count/sample index is a good enough proxy for a measure of practice, noting that not all practice is equal [11] and timing/spacing of instances of practice significantly affect skill development [22].

A single biometric analysis distributes a selection of samples across a training and test set and reports on results. In order to account for practice, it is important to look at how many and which samples are discarded (not used for

ISBN 978-90-827970-1-5 © EURASIP 2018

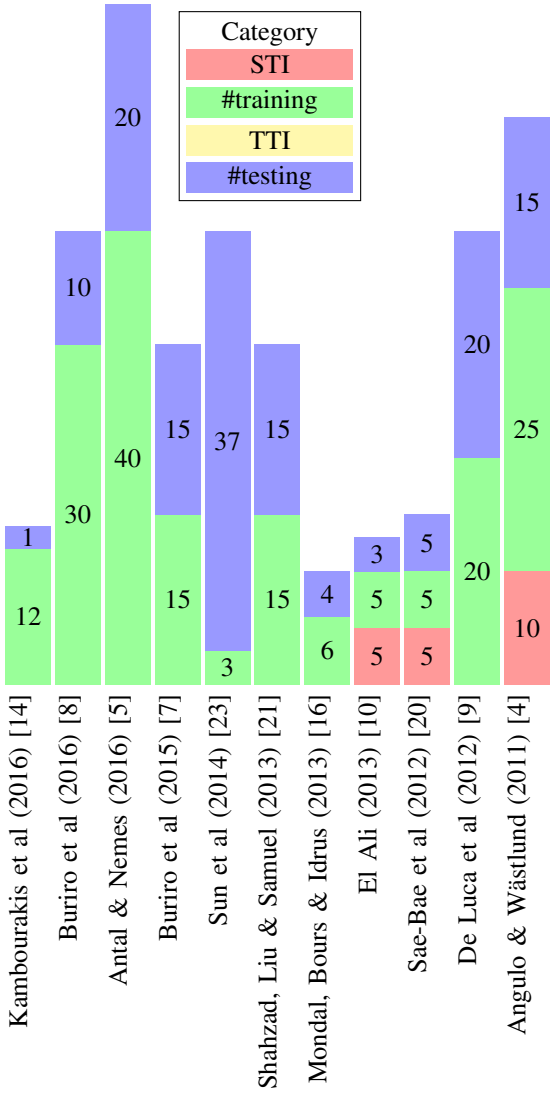

Figure 1. An overview of how samples fell into tutorial, training, usage and testing categories per participant for different recent studies.

the analysis), both those before the training samples, and before the test samples. We dub these sample intervals the Start-to-Train Interval (STI) and Train-to-Test Interval (TTI) respectively. Increasing the STI makes us discard more initial samples, allowing us to look at how different levels of initial skill affect recognition performance. Similarly, increasing the TTI simulates extended use of the behavioural biometric application, and how recognition performance develops with use. Figure 1 shows an overview of how samples were picked for a selection of recent studies; note how only some used (in our terms) an STI, and none used a TTI, indicating that there is an incognizance in literature when it comes to this topic.

Again, it is unclear whether the hypothesized effects, that with practice recognition performance both increases and stays stable for longer, really exist and whether these effects are strong enough to matter. However, from the fact that there is rich literature on practice in motor skill learning, and that it is something under-investigated in contemporary behavioural biometric literature, we can make the case that these effects are likely significant and worth investigating for any new behavioural biometric application. In using the concepts of sample order intervals, the STI and TTI, we can analyze data systematically with regards to practice, ideally allowing for higher comparability of study results [15].

The aim of this paper is therefore twofold. First, we use the STI/TTI concept to introduce a novel method of presenting recognition performance with regards to practice. Two, we use this method to analyze three existing data sets from motor skill studies, in which high amounts of repetition (500+) is the norm, to investigate the existence of our hypothesized effects of practice on biometric recognition performance, and report on our findings.

II. Methods

A. Data sets

Table II shows an overview of the three data sets we analyzed. All data sets were gathered through a task paradigm called the Discrete Sequence Production (DSP) task. A DSP task is basically a repeated entry of a very simple password, see [1] for details. All three data sets included the practice of 2 individual sequences/passwords. For both [25] and [26], these were 6-character passwords. For [6], there was one 6-character and one 3-character password. We did not analyze the 3-character password. A caveat of these data sets is that no key-up times were recorded, only key-down, meaning recognition performance is likely lower than it could have been. We therefore report on trends in relative recognition performance rates, rather than interpreting the the absolute values.

Name	# Participants	# Repetitions	# Blocks	# Repetitions/Block	Break (min)
Verwey2009a [25]	48	518	7	74	7
Verwey2009b [25]	48	518	7	74	7
Verwey2016a [26]	24	600	15	40	4
Verwey2016b [26]	24	600	15	40	4
Barnhoorn2017a [6]	32	432	18	24	2

Table II
Overview of the analyzed data sets.

B. Analyses

For our analyses, we only considered "practice" blocks, in which more straight-forward password entry behaviour is requested of participants. Between blocks, breaks were included, we refer to [6], [25], [26] for specifics on how these breaks were structured. Our samples consisted of 7 features; 6 key-down values, as well as a precise accuracy value indicating if and where mistakes were made. We imputed missing values by a unrealistically high key-down time (1e6 seconds). We used the out-of-the-box random forests (RF) classifier available in the python library scikit-learn [17], only default settings were chosen. RF classifiers are resistant to

ISBN 978-90-827970-1-5 © EURASIP 2018

outliers, such as those introduced by our imputation strategy. They also have the added benefit of being somewhat introspectable, as they are based on decision tree classifiers, which is an interesting property to have moving forward with similar studies. We report the recognition performance of our classifier in terms of Equal Error Rates (EER).

We chose a sample size of 24 for both the classifier training and test sets. We incremented the STI and TTI in steps of 12 and exhaustively tested all combinations of STI and TTI against eachother. As our intervals are bounded by the amount of samples in the data set, the STI + TTI is always less than the total amount of samples per sequence.

For representation of the EERs, we introduce a novel presentation method we named the STI/TTI contour plot, in which we plot the EER against the STI and TTI values that generated it. This creates a plot filling the lower triangle of the graph, bounded by the line STI + TTI = #samples. The STI/TTI contour plot was chosen specifically because it makes it easy to see how recognition performance develops under practice qualitatively.

III. Results

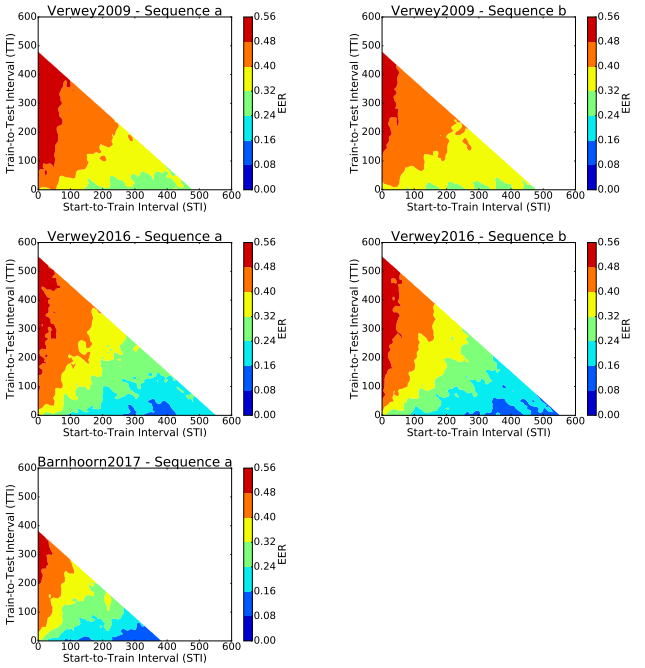

Table III
STI/TTI CONTOUR PLOTS OF THE ANALYZED SEQUENCES.

We present the individual STI/TTI contour plots in Table III. The most salient feature of these contour plots is how the areas of EER seem to fan out from the origin point (where STI == TTI == 0). These patterns hold for all contour plots, but we generally find lower EERs for both verwey2009 sequences, with a minimum in the 0.24-0.32 EER range, than for the other sequences, with a minimum EER in the 0.08-0.16 range. There seem to be some interesting artifacts that make the contours look jagged, and although some are probably due to the hard thresholding between values of EERs that are likely very similar (very noticeable in verwey2016a), others can be explained by the (length of) breaks between repetition blocks [25], which we will expand on later.

Looking at the Verwey2009 and Verwey2016 studies, we see that the specific form of the contour plots differ quite a bit between the two studies. However, within these studies the contour plots are highly similar. The shape of the contour plots suggest some relationship between the STI/TTI and the EER, a relationship apparently influenced by the specific set-up of the individual studies.

From the individual contour plots, we can see two general trends. First, as we increase the STI, we find that our recognition performance generally becomes lower (barring some exceptions). This suggests a more stable recognition performance as we allow for more initial practice. As we increase the TTI however, we find that recognition quickly deteriorates, at times to the point where the decision is effectively random (an EER of .5).

There also seems to be some interaction between STI and TTIs. If we take vertical slices at increasing STI values, we find that our EER deteriorates with increased TTIs, but we also find that this rate of deterioration decreases with an increased STI. Suggesting that the more participants are allowed to practice, the longer the recognition performance stays stable. If we take horizontal slices of the contour plots at increasing TTIs, we simulate how the recognition performance of specific applications will be at different points in their lifetime, dependent on how much practice a user was allowed before enrollment. As we increase the STI, the EER becomes better, but we also find that with a lower TTI, the rate of EER change is higher.

Some of the earlier mentioned artifacts can be explained by experimental design decisions. There is a sine-wave-like periodicity in the EER in the Verwey2009 sequences, most noticeable at low TTIs. The half of the periodicity is about 70 samples long, exactly the amount of repetitions in a single practice block [25]. A similar dip in EER can be seen in barnhoorn2017a around the 225 sample mark, which coincides with a break between day 1 and day 2. Further, we note that the magnitude of these artifacts is similar, although breaks were hugely different in time (7 minutes vs 1 day).

IV. Discussion

Our primary goal in this paper was to investigate how practice influences recognition performance and its stability. Specifically, we hypothesized that more accumulated practice both results in increased and more stable recognition performance. Both seem supported by the DSP data sets we analyzed, as can be fairly easily seen from the STI/TTI contour plots.

Increased recognition performance with more practice is the case across the STI axis. For all values of TTI, barring some artifacts in the contours, we can see much lower EERs as one increases the STI. Similar increases in recognition performance can be seen across sequences within a study, an indication that whatever this specific effect of practice is, it is stable within participants across multiple instances of the

ISBN 978-90-827970-1-5 © EURASIP 2018

same type of sequence. We can also see increased stability of recognition performance with more practice is generally the case from looking at the rate at which it deteriorates. Our recognition performance seems to always deteriorate as one increases the TTI. However, this rate is much slower for high STI values than for low STI values. This suggests that the behavioural patterns our classifier learned to recognize are unstable, but less unstable for skilled than for unskilled individuals.

From a psychological perspective, it is unsurprising that behaviour continues to change for a long time in ever slower steps [2]. However, as is also shown in [2], different types of correlations (features) become more and less predictive of underlying theoretical structures in different stages of practice (novices, intermediate, expert). If the strength of these theoretical structures correlate with recognition performance, it becomes possible that our deterioration in recognition performance is confounded by our specific features losing their predictiveness with more practice. There might also be different features that do not lose their predictiveness and thus are more resilient to the effects of practice on recognition performance. This could be investigated in current data sets by looking at the types of features that white-box classifiers such as Random Forests find most discriminating, and seeing if/how these change under practice, however this investigation is limited by the types of features we have at our disposal and would ideally require collection of new data sets.

Another caveat that needs mentioning is that in these studies, many repetitions are performed in a short period of time, whereas in a typical use-case for a behavioural biometric application repetitions will be far more spread out in time. Given the same environments, the typical use-case thus has to deal with a lack of task focus, whereas small numbers of longer sessions are more likely to deal with hyper-focus and fatigue. How these effects compare is unknown in behavioural biometric literature [19], but further investigations in practice effects should give us more insight into how such comparisons can be made.

Further, we did not vary the sample size of our training or test sets. It is possible that larger sample sizes would make our recognition performance more robust to changes in behaviour, as the classifier is able to learn from a wider range of behavioural patterns. On the other hand, a smaller sample size, as also reported by [4], could possibly have increased performance, especially as rates of behavioural change are very high in the early, low practice phases of the experiment. It should be clear however, that one might choose different samples sizes depending on the average skill the population of users of a behavioural biometric application are likely to have. Similarly, if one can learn to estimate skill levels, individually tailoring enrollment phases to individuals should be possible.

We also reported on some miscellaneous findings. Breaks between practice blocks seem to introduce artifacts, locally lowered EERs, at points where these breaks happened. Short breaks of 2 minutes introduced no discernible artifacts, but longer breaks of 7 minutes and in one case a full day seemed to introduce artifacts of similar magnitude. This suggests non-linear effects of time on the stability of our behavioural patterns, and thus our recognition performance. Although the effects of practice spacing, as it is known, as well as sleep on skill improvement is heavily researched (e.g. [22]), not much is known of its effects on the stability of our behaviour. This might present an interesting opportunity for collaboration between the fields of behavioural biometrics and cognitive psychology.

A final remark has to be made regarding the specific experimental designs of the studies we looked at. In general, in behavioural experiments one variable is identified and systematically changed to see what effects (or lack of effects) this variable will have on our behaviour. Oftentimes, this variable is a demographic one, for example in the case of [6], where the differences in behaviour between groups of older adults (avg. age 79) and young adults (avg. age 21) was tested. For this manuscript, we disregarded such variables, as we wanted to focus on conceptually introducing a toolset to learn to understand the effects of practice. However, what the effects of such demographic factors in combination with practice are on recognition performance should definitely be investigated at some point. A reading of [6] suggests effects of age might be especially pronounced for low practice, but dissappear after high amounts of practice.

A. Conclusion

In this paper we set out to better understand the effects of practice on behavioural biometric recognition performance. We found that, as hypothesized, recognition performance increases and becomes more stable as individuals accumulate more practice and become more skilled. However, this requires significantly more practice than what is usually allowed in behavioural biometric validation studies. These results are possibly confounded by several factors, including available features and highly packed repetitions in time. We will therefore follow up on this with the collection of more data sets in order to account for said shortcomings.

Regardless, we believe the STI/TTI contour plot is a powerful novel method to systematically analyse effects of practice and show that these effects are clearly significant. As such, we believe this to be a step in the right direction towards understanding the effects of practice on recognition performance, in order to build behavioural biometric applications that are more robust to it.

V. Acknowledgements

This study was funded by Ubiqu Access B.V. Author E. Haasnoot is supported by the aforementioned company and declares he has no conflict of interest. DSP task data collection was supported previously by several grants, information on which can be found in the related papers [6], [25], [26]. We thank Bram Kolkman for assistance in designing several figures.

References

[1] Abrahamse, E. L., Ruitenberg, M. F., De Kleine, E., and Verwey, W. B. Control of automated behavior: insights from the discrete sequence production task.

[2] ACUNA, D. E., WYMBS, N. F., REYNOLDS, C. A., PICARD, N., TURNER, R. S., STRICK, P. L., GRAFTON, S. T., AND KORDING, K. P. Multifaceted aspects of chunking enable robust algorithms. *Journal of neurophysiology 112*, 8 (2014), 1849–1856.

[3] ALBRECHT, S., JANSSEN, D., QUARZ, E., NEWELL, K. M., AND SCHÖLLHORN, W. I. Individuality of movements in music–finger and body movements during playing of the flute. *Human movement science 35* (2014), 131–144.

[4] ANGULO, J., AND WÄSTLUND, E. Exploring touch-screen biometrics for user identification on smart phones. In *IFIP PrimeLife International Summer School on Privacy and Identity Management for Life* (2011), Springer, pp. 130–143.

[5] ANTAL, M., AND NEMES, L. The mobikey keystroke dynamics password database: Benchmark results. In *Software Engineering Perspectives and Application in Intelligent Systems*. Springer, 2016, pp. 35–46.

[6] BARNHOORN, J., VAN ASSELDONK, E., AND VERWEY, W. Differences in chunking behavior between young and older adults diminish with extended practice. *Psychological research* (2017), 1–11.

[7] BURIRO, A., CRISPO, B., DEL FRARI, F., AND WRONA, K. Touchstroke: smartphone user authentication based on touch-typing biometrics. In *International Conference on Image Analysis and Processing* (2015), Springer, pp. 27–34.

[8] BURIRO, A., CRISPO, B., DELFRARI, F., AND WRONA, K. Hold and sign: A novel behavioral biometrics for smartphone user authentication. In *Security and Privacy Workshops (SPW), 2016 IEEE* (2016), IEEE, pp. 276–285.

[9] DE LUCA, A., HANG, A., BRUDY, F., LINDNER, C., AND HUSSMANN, H. Touch me once and i know it's you!: implicit authentication based on touch screen patterns. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2012), ACM, pp. 987–996.

[10] EL ALI, A., ET AL. *Minimal mobile human computer interaction*. Universiteit van Amsterdam [Host], 2013.

[11] ERICSSON, K. A., KRAMPE, R. T., AND TESCH-RÖMER, C. The role of deliberate practice in the acquisition of expert performance. *Psychological review 100*, 3 (1993), 363.

[12] GAINES, R. S., LISOWSKI, W., PRESS, S. J., AND SHAPIRO, N. Authentication by keystroke timing: Some preliminary results. Tech. rep., DTIC Document, 1980.

[13] HEITZ, R. P. The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Frontiers in neuroscience 8* (2014), 150.

[14] KAMBOURAKIS, G., DAMOPOULOS, D., PAPAMARTZIVANOS, D., AND PAVLIDAKIS, E. Introducing touchstroke: keystroke-based authentication system for smartphones. *Security and Communication Networks 9*, 6 (2016), 542–554.

[15] KILLOURHY, K. S., AND MAXION, R. A. Comparing anomaly-detection algorithms for keystroke dynamics. In *Dependable Systems & Networks, 2009. DSN'09. IEEE/IFIP international conference on* (2009), IEEE, pp. 125–134.

[16] MONDAL, S., BOURS, P., AND IDRUS, S. Complexity measurement of a password for keystroke dynamics: Preliminary study. In *Proceedings of the 6th International Conference on Security of Information and Networks* (2013), ACM, pp. 301–305.

[17] PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M., AND DUCHESNAY, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[18] REIS, J., SCHAMBRA, H. M., COHEN, L. G., BUCH, E. R., FRITSCH, B., ZARAHN, E., CELNIK, P. A., AND KRAKAUER, J. W. Noninvasive cortical stimulation enhances motor skill acquisition over multiple days through an effect on consolidation. *Proceedings of the National Academy of Sciences 106*, 5 (2009), 1590–1595.

[19] REVETT, K. *Behavioral biometrics: a remote access approach*. John Wiley & Sons, 2008.

[20] SAE-BAE, N., AHMED, K., ISBISTER, K., AND MEMON, N. Biometric-rich gestures: a novel approach to authentication on multi-touch devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2012), ACM, pp. 977–986.

[21] SHAHZAD, M., LIU, A. X., AND SAMUEL, A. Secure unlocking of mobile touch screen devices by simple gestures: you can see it but you can not do it. In *Proceedings of the 19th annual international conference on Mobile computing & networking* (2013), ACM, pp. 39–50.

[22] STAFFORD, T., AND HAASNOOT, E. Testing sleep consolidation in skill learning: A field study using an online game. *Topics in cognitive science 9*, 2 (2017), 485–496.

[23] SUN, J., ZHANG, R., ZHANG, J., AND ZHANG, Y. Touchin: Sightless two-factor authentication on multi-touch mobile devices. In *Communications and Network Security (CNS), 2014 IEEE Conference on* (2014), IEEE, pp. 436–444.

[24] UMPHRESS, D., AND WILLIAMS, G. Identity verification through keyboard characteristics. *International journal of man-machine studies 23*, 3 (1985), 263–273.

[25] VERWEY, W. B., ABRAHAMSE, E. L., AND JIMÉNEZ, L. Segmentation of short keying sequences does not spontaneously transfer to other sequences. *Human movement science 28*, 3 (2009), 348–361.

[26] VERWEY, W. B., GROEN, E. C., AND WRIGHT, D. L. The stuff that motor chunks are made of: Spatial instead of motor representations? *Experimental brain research 234*, 2 (2016), 353–366.

[27] VIVIANI, P., AND LAISSARD, G. Motor templates in typing. *Journal of Experimental Psychology: Human Perception and Performance 22*, 2 (1996), 417.

ISBN 978-90-827970-1-5 © EURASIP 2018