ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/317400387

Predicting Age and Gender by Keystroke Dynamics and Mouse Patterns

Conference Paper · July 2017

DOI: 10.1145/3099023.3099105

CITATIONS
5

READS
5,540

1 author:

Avar Pentel
Tallinn University
10 PUBLICATIONS **24** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Mining Unintentional Traces of Computer Usage View project

All content following this page was uploaded by Avar Pentel on 08 November 2017.

The user has requested enhancement of the downloaded file.

Predicting Age and Gender by Keystroke Dynamics and Mouse Patterns

Avar Pentel
Tallinn University, School of Digital Technologies
Tallinn, Estonia
pentel@tlu.ee

ABSTRACT

In human computer interaction, some of the user activities are intentional, and other unintentional, but user interfaces are usually designed to react only to intentional commands. However, user's unintentional activity contains many clues about a user, that can be beneficial to take into account in designing appropriate response. Current study focuses on these unintentional traces, that left behind by use of standard input devices, keyboard and mouse, and specifically, we try to predict users age and gender. Mouse and keyboard data used in this study, are collected in six different systems between 2011 and 2017 in total from 1519 subjects. Some supervised machine learning models yield to f-scores over 0.9 when predicted both user age or gender.

CCS CONCEPTS

• **Computing methodologies** → **Instance-based learning**; *Supervised learning by classification*;

KEYWORDS

User modeling, age prediction, gender prediction, keystroke dynamics, mouse dynamics

ACM Reference format:
Avar Pentel. 2017. Predicting Age and Gender by Keystroke Dynamics and Mouse Patterns. In *Proceedings of UMAP'17 Adjunct, Bratislava, Slovakia, July 09-12, 2017,* 6 pages.
https://doi.org/10.1145/3099023.3099105

1 INTRODUCTION

Knowing anonymous user age and gender can be beneficial in many areas, for instance, in recommendation systems in online marketing or in detecting cyber criminals, pedophiles or preventing underaged viewing some inappropriate content. There are many works done in gender and age detection by analyzing the content, i. e. text's written by anonymous authors [6, 9, 12, 13], for example. The problem of this approach is the length of the text what is needed from anonymous author in order to make predictions. The more text we have, the more precise predictions we can make. In previous works, we tried to make classification on very short texts using stylometric features [12] with n-gram and n-graph features [13]. Giving full credit to text mining approaches, they cannot cover cases where users are interacting with online system without writing anything. Therefore, we focus on the current study on less studied area - namely user modeling by keystroke dynamics and mouse movement patterns.

The idea behind keystroke dynamics is known since World War II. Telegraph operators could recognize the sending operator because of the uniqueness in the keying rhythm [18]. Later, similar methods were used to analyze computer keyboard usage. Now, keyboard dynamics have been studied over 30 years [5, 19], but mostly in connection to authentication. Some of those identification studies [11], reported different factors that affected the identification results. Among these factors were also age, gender, and experience. This gives some ground to hope that filtering out individual differences, we can find some common features that distinguish between user age or gender.

Effect of age and experience on typing are reported already back in 1984 [17] in psychological literature. User profiling by keystroke dynamics is a more recent field of study. Some related studies employed keystroke dynamics on gender prediction, and yield to 80% accuracy [4] and to 91.63% accuracy [8]. In another study models were generated for both age and gender [2], which identified user under the age of 15 with up to 91%. Another study [1] collected data from large number of subjects, but the classification results were far more modest. User modeling based on mouse movement patterns is even less studied. There are still several works on authentication, emotion detection and some works on detection of age [3] and gender [10].

We suppose that when experience influences keystroke dynamics, then the same features that distinguish classes in text mining, can be employed as features in keystroke dynamics. For example, when in text mining as features character n-graph frequencies are used and some character n-graph characterize better female or male, young or old group, then it probably means that, they used to type these n-graphs more often and are more experienced to do so.

It seems intuitively plausible, that this experience reflects in typing speed relative to other n-graphs. Unfortunately, our datasets for keystroke dynamic data are all in Estonian language, and there are only a few text mining studies [12, 13] based on the Estonian language and they classified only two age groups - younger group 7-15 versus older group. However, we will test in the current study if those frequency based text mining features have similar influence on typing rhythm. Secondly, we will test if there are significant differences in typing speed between different age groups and gender groups. And finally we create machine learning models to predict age and gender.

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.
UMAP'17 Adjunct, July 09-12, 2017, Bratislava, Slovakia
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-5067-9/17/07…$15.00
https://doi.org/10.1145/3099023.3099105

Table 1: Sample.

	Description	number of subjects total	female	male	number of instances keystroke	mouse
1.	K12 School Intranet (collected 2011- 2017)	67	54	13	64	n/a
2.	Feedback from students, parents, and employees (2014-2017)	554	333	221	554	554
3.	Emotion induction and detection experiments [15]	64	35	29	256	1536
4.	Test of aesthetic preferences (High school project 2016)	172	62	110	172	172
5.	Personality and handedness test (High School project 2017)	485	388	97	485	485
6.	User attention test (High School project 2017)	177	125	52	177	2301
	Total:	1519	997	522	1708	5048

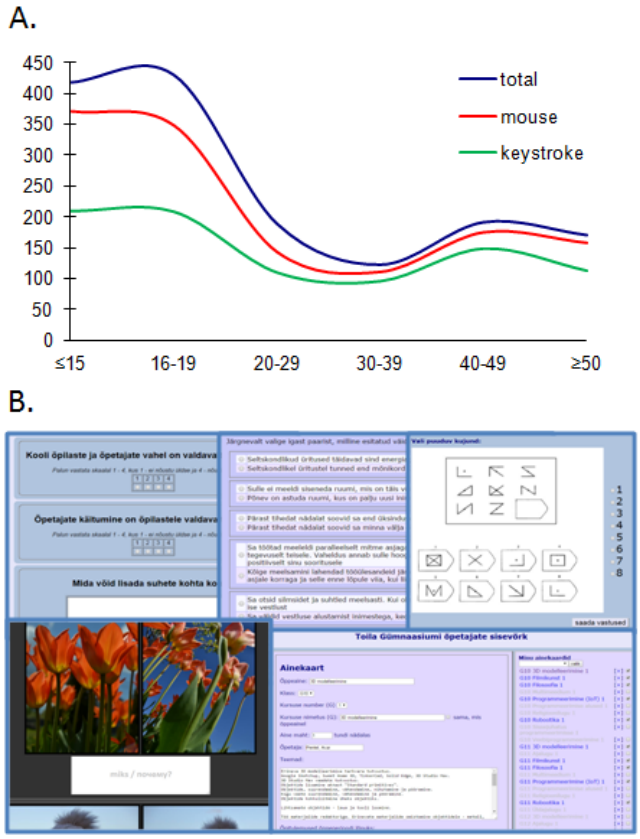

Figure 1: Distribution of age groups (A) in dataset and different contexts (B) of data collection. Screen shots of data sources 2, 5, 3, 4, and 1).

If compared to previous studies, our keystroke and mouse datasets are larger, the number of individual subjects is higher, and the data are collected from very different sources. In many cases, data collection was carried out in a natural environment, while the users were performing their everyday activities related to their work or study. This ecological validity minimizes environmental influences, and makes our results more generalizable.

The rest of the paper is structured as follows. In Section 2. we give an overview of our data sources, data collection methods, feature extraction and data analysis methods. In section 3. we present the results and in section 4 we conclude this paper.

2 EXPERIMENTAL DETAILS

2.1 Overview of experiments and samples

The current study is based on data we collected in six sources between 2011-2017 (Table 1). Two of the sources were real life working systems, the intranet and on the feedback questionnaire of the Estonian k12 school of Toila. Age based distribution of our dataset as well as the different contexts of data collection environments is illustrated in Fig. 1. The reason why there are different distributions for mouse and keystroke data is because in many cases the keyboard input was not mandatory and as a result, we had a number of empty logs, that are excluded here. We also excluded logs that contained only a few keystrokes. Similarly the we excluded from mouse logs these created by mobile or other touch screen devices.

In first system school personnel can input different kind of data, filling lesson plans, uploading or creating learning materials, writing events to the calendars, writing reports, different notifications to the public web, etc. Partly it was as the content management system for public web page, partly closed system for employees, and students, and partly private. This data collection began in 2011, and continues today. In this system only keystroke dynamics data are logged. The second data source - feedback questionnaire - has 72 multiple choice and 8 free answer questions, and was administered to all pupils from 5th to 12th grade, for all parents and for all employees. For each group is slightly different questionnaire, and previously mentioned number of items is from pupil questionnaire.

In these online questionnaires keystroke data and mouse logs have been collected since 2014. Other three data sources were actually experiments, but only one of them was specifically designed for studying keystroke dynamics and mouse patterns (Table 1. 3.), and the main focus there was emotion detection [15]. The rest of the three where high school project questionnaires, 2015-2016, and 2016-2017, accordingly, where the aim of these projects was not related to studying keystroke and mouse dynamics, but we cooperated with these teams in order to use these questionnaires as a platform for our own data collection.

As we can see in Table 1, the number of instances is different for keyboard and mouse. It is because in data source 3, there was only one typing task for each of four experiments, but 6 tasks for a mouse.

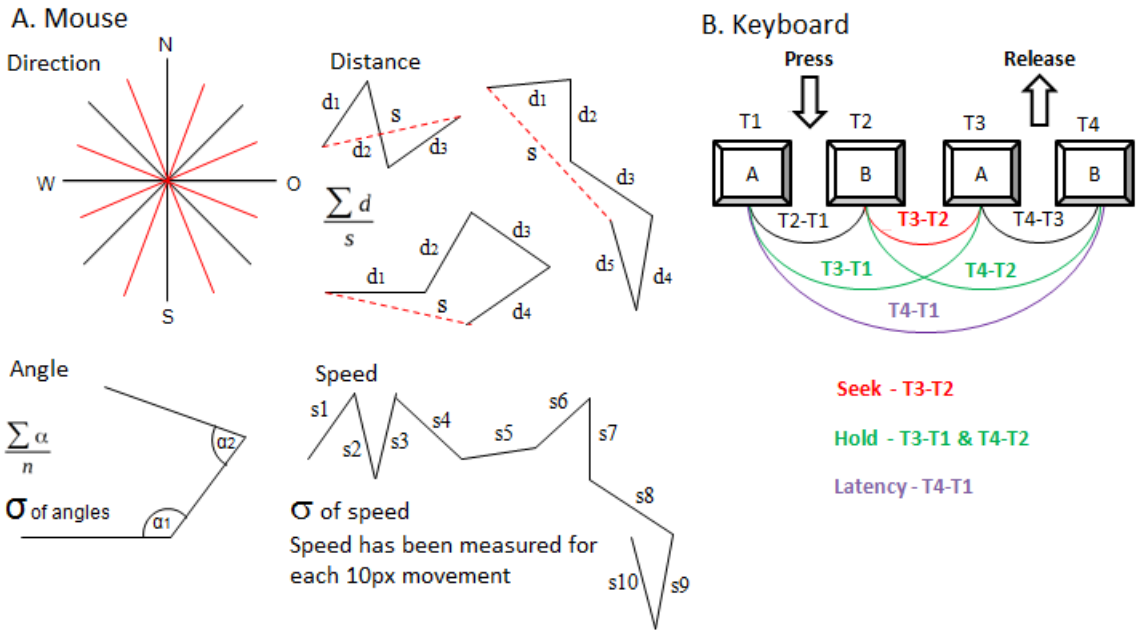

Figure 2: Types of mouse movement and keystroke features

In data source 6 the test was designed on page-to-page manner and mouse logs were saved for each page separately. Because of few input elements keyboard data were collected as one log per user.

It is worth of mentioning that the amount of typing data, and representativeness varies by sources. So in the first data source there is a large amount of keyboard data, but it is created by a small number of people, mostly female and all adults over 25 years old. Then we have data from relatively long feedback questionnaires, where all pupils were presented and thus presenting different age groups from 11-19 years and females and males are evenly distributed.

On all of those questionnaires the most of free text input fields were not required to be filled, and the length of typed text was relatively short. All except one data source presented free text, only data source 3 presented semi fixed text, where users had to type three verses of Estonian national anthem from memory.

Only a small part of our data were collected in a controlled environment, namely data source 3 in Table 1 and part of data source 2 which concerns pupils. In order to still get some overview about used technology, our system recorded users screen resolution and user agent string that contains information about used web browser, operating system, and if the used device was a computer, or some mobile device. However, in this study we included all the data in keystroke data collection not depending on if it was typed in using a physical keyboard or on a virtual keyboard on touchscreen devices.

2.2 Data collection technology

2.2.1 Mouse Data. Web based mouse movement logging system was created using JavaScript exactly in the same way as in previous studies [14]. In this system the mouse positions are not logged in constant interval, but by constant distance. This means that the system is constantly monitoring the mouse pointer position and when there is the difference over 10 pixels from initial position then new position will be logged as x, and y coordinates and the timestamp. Thereafter the system will measure distance from this new point. This means that in our case the logged mouse path is simplified to 10 pixel length straight lines. In practice, and taking into account the mouse polling time, in case of fast mouse movements, the real distance between two logged points might be bigger.

2.2.2 Keyboard Data. Keystroke data were logged using JavaScript based key logging system described in [12]. The script registers *keydown* and *keyup* events and timestamps, and based on this information calculates and logs three numbers for each key press:

(1) *keyCode*, which is a numeric representation of a key,
(2) the time when a key was pressed, and
(3) the time when the key was released.

2.3 Feature Extraction

2.3.1 Mouse features. We used four types of mouse movement features from previous study [14]. These features were based on direction, angle, distance, and the velocity - Fig. 2. A. Direction based features were calculated as the relative number of movements in a particular direction. Distance based features were calculated as the ratio of full paths to shortest path. For each 10 pixel length movement pair the angle was calculated between them and average angle and standard deviation of angles was used as a feature. For each 10 pixel length movement the velocity was calculated and the standard deviation of velocity was used as a feature.

Table 2: Extracted keystroke features

No	Feature type	Total
1.	Hold time	59
2.	Seek time	59
3.	n-graph latency (n: 2-4)	335, 418, 172
4.	Feature Std for each feature	1043
5.	DEL	2
6.	Correct	1
7.	Time	1

2.3.2 Keyboard features. As there are different terminologies in use, relating to keystroke feature types, we define the most important terms how we use them in the current study. Our main keystroke feature types were calculated as hold times, seek times, and n-graph latencies as in Fig. 2. B. By hold time we mean time between key pressed down and released. Seek time is time between last key is released and a new key is pressed.

Combinations of keys are characterized by n-graph latencies, which means the time it takes to type particular n number of keys starting from first key press and ending with the last key release. All used feature types are defined in the following list:

(1) Hold time - mean time between key press and release for particular key.
(2) Seek time - mean time between last key release and current key press for particular key.
(3) n-graph latency - mean time of n consecutive key presses, starting from first key press and ending with the second key release for particular n-graph.
(4) Feature Std - standard deviation for each of the previous features.
(5) DEL - relative frequency of corrective keys.
(6) Correct - number of keystrokes divided to number of characters in final text
(7) Time - mean time of all typed keys (seek+hold) for a particular user.

The total number of extracted features for each instance is given in Table 2. All time-based keystroke features (Table 2, 1 - 3) were standardized for each subject, in order to minimize individual differences in typing speed. The possible number of features based on seek time, hold and n-graph latencies is in reality bigger, here we limited all those features by the frequency in full dataset and used only those features that were presented more than 1000 times. Chi-square attribute evaluation was carried out before generating models, which again reduced the number of features. Missing values were replaced by feature mean.

2.4 Models

2.4.1 Machine learning and technology. For modeling, we tested five popular machine-learning algorithms for binary classification: Logistic regression, Support Vector Machine, Nearest Neighbor, C4.5, and Random Forest. In our task we used Java implementations of listed algorithms that are available in freeware data analysis package Weka [7]. Before generating models, we balanced our datasets under-sampling the majority class. Used data had gender based labels and age group labels for 10-15 years old, 16-19, 20-29, 30-39, 40-49, 50+, but in the current study we represent only classification of 10-15 years old against all others.

There are three reasons for doing such binary classification between only two classes. First, while our under-aged group is the largest, it allows to balance groups with less under-sampling. Secondly, we want to compare our best time-based n-graph features with frequency based n-graph features that are reported in previous text-mining study, and therefore we follow the same group division that was made there. And the third reason is more juridical, it is based on the age of consent. By Estonian law the age of consent is set to 14 years, it is subject of political discussion, and most probably change will be upward 2 years. If utilizing our results in the prevention of criminal activities, cases like child sexual abuse or pedophilia, this particular distinction between age classes is also justified.

However, there are many utilities to distinction between all age groups, and in follow up study [16], we classified all other age groups too. We do the first step in that direction here by comparing mean typing time between all the age groups.

2.4.2 Evaluation. For evaluation, we used 10 fold cross validation on all models. It means, we partitioned our data into 10 even sized and random parts, and then using one part for validation and another 9 as training dataset. We did so 10 times and then averaged validation results. Presenting our results, we use a single f-score value, which is a weighted average of both classes' f-score values.

2.5 Comparing Group Means

To test if there are significant differences between groups in typing speed we also preformed two sample T-tests between each age group. Here we also compared all other age groups. Similarly, we compared mean typing speed between male and female subjects.

2.6 Comparing Keystroke and Text Mining Features

We compared top 100 keystroke n-graph features with text mining features from previous study [16]. When in keystroke dynamics the n-graph feature represents the time, then in text-mining n-graph feature represents frequency. We searched for overlap between sets of top-ranked features for both studies.

3 RESULTS AND DISCUSSION

The results of the gender and age based classification are presented in Table 3. Baseline for each classification was 0.5. While all results are over the baseline, here we see better results with models trained on mouse features. This might be partly because we had more mouse instances from the same users and for instance in case of Nearest Neighbor with k=1, we cannot be sure if the result shows the quality of classification or identification. At least this phenomenon needs further investigation.

If investigating attribute weights on models and attribute selection results, then it seems that most important keystroke predictors of age are related to variations in typing speed. Similarly, some of the standard deviation based mouse features played role in age

Table 3: F-scores of different models and classification tasks

Model	f-scores			
	Males vs Females		10-15 vs 16+	
	keyboard	mouse	keyboard	mouse
Logistic Regression	0.69	0.69	0.6	0.86
SVM (std)	0.73	0.69	0.66	0.86
kNN (k=1)	0.73	0.94	0.6	0.95
C4.5	0.67	0.81	0.65	0.86
Random Forest	0.73	0.88	0.73	0.92

Table 4: T-test results (two-tailed). Comparing typing speed

Age	16-19	20-29	30-39	40-49	50+
10-15	9.88***	5.54***	-0.71	-0.3	2.93*
16-19		-1.92	-7.74***	-7.74***	-9.79***
20-29			-4.47**	-4.4**	-6.02***
30-39				0.34	-1.54
40-49					-2.08*
			*p < 0.05,	**p<0.001,	***p < 0.0001

based predictions. Between females and males there were no dominantly important distinguishing features or feature types. Among other keystroke features, corrective keys, average time of all keys and features based on standard deviation had a bigger info gain.

Interesting finding related to age is that from time based features, i.e., hold and seek time and n-graph latencies, out of top twenty features, 11 were digraphs and trigraphs containing spaces in the end or in the beginning. It seems that the pauses between words itself are good predictors of age and it is reasonable in further studies to create one combined feature that measures the average time of all n-graphs with space at one end.

From single key features seek time was a better predictor of age than hold time. Among mouse features the biggest info gain was related to all distance-based features, standard deviation of velocity and standard deviation of angles.

T-test results did not reveal any significant differences between females and males, however, between the age groups, some significant differences emerged, as seen in Table 4. It seems that the age group 16-29 is significantly faster in typing than other groups. Which is intuitively understandable, because of generational differences. The youngest group, in other hand, is still gaining the experience or is their slower typing related to use of different kind of mediums.

We compared top n-graph frequency based features with top features from the text mining study. 22 out of top 100 ranked n-graph features had overlap with top 100 features from previous text mining study [12], which supports our initial assumption that the same features that distinguish age groups in text mining are more familiar to group members and that reflects in typing speed. We had no keystroke data from this previous text mining study and similarly we did not text mining on current datasets, therefore we plan to investigate this relationship on further studies when we can use both text-mining and keystroke datasets from the same input results from the same dataset.

4 CONCLUSIONS

We have collected keystroke and mouse dynamic data from six different sources. Combining this data together we performed user modeling in order to predict anonymous user age and gender. All models predicted better than chance. The diversity of our data sources gives ecological validity to our results.

We also found significant differences in typing speed between different age groups. Comparing our best time based features with frequency based features from previous studies, we also found some overlap between the two sets, which emphasizes the relationship between text familiarity and typing speed.

The main contribution of this study, is on minimization context dependency of our results by collection of large sets of keystroke and mouse data from very diverse sources. The results of this study are still preliminary, and we are doing follow-up studies in order to improve our classification results and to get a better understanding what are the general links between age and gender and keyboard and mouse usage.

REFERENCES

[1] D. G. Brizan, A. Goodkindand, P. Koch, K. Balagani, V. V. Phoha, and A. Rosenberg. 2015. Utilizing linguistically enhanced keystroke dynamics to predict typist cognition and demographics. *International Journal of Human-Computer Studies* 82 (2015), 57–68.

[2] D. Chuda and P. Kratky. 2015. Grouping Instances in kNN for Classification Based on Computer Mouse Features. In *16th International Conference on Computer Systems and Technologies*.

[3] A. Dantcheva, C. Velardo, A. D' Angelo, , and J.-L. Dugelay. 2011. Bag of soft biometrics for person identification. *Multimedia Tools and Applications* 51, 2 (2011), 739–777.

[4] M. Fairhurst and M. D. Costa-Abreu. 2011. Using keystroke dynamics for gender identification in social network environment. In *4th International Conference on Imaging for Crime Detection and Prevention (ICDP 2011)*.

[5] J. Garcia. 1986. Personal identification apparatus. *US Patent Office* 4621334 (1986).

[6] S. Goswami and S. S. M. Rustagi. 2009. Stylometric analysis of bloggersâĂŹ age and gender. In *Third International AAAI Conference on Weblogs and Social Media (ICWSM 2009)*.

[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The WEKA data mining software: an update. *SIGKDD Explorations* 11, 1 (2009).

[8] S. Z. S. Idrus, E. Cherrier, C. Rosenberger, and P. Bours. 2014. Soft biometrics for keystroke dynamics: Profiling individuals while typing passwords. *Computers & Security* 45 (2014), 147âĂŞ155.

[9] M. Koppel, S. Argamon, and A. R. Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing* 17, 4 (2002), 401–412.

[10] P. Kratky and D. Chuda. 2016. Estimating Gender and Age of Web Page Visitors from the Way They Use Their Mouse. In *WWW '16 Companion Proceedings of the 25th International Conference Companion on World Wide Web*.

[11] F. Monrose and A. D. Rubin. 2000. Keystroke Dynamics as a Biometric for Authentication. *Future Generation Computer Systems. Elsevier* 16 (2000), 351–359.

[12] A. Pentel. 2015. Effect of different feature types on age based classification of short texts. In *Intelligence, Systems and Applications (IISA2015)*. IEEE Digital Library.

[13] A. Pentel. 2015. Employing Relation Between Reading and Writing Skills on Age Based Categorization of Short Estonian Texts. In *In Proceedings of DeCAT 2015*, L. Aroyo, G. Houben, P. Lops, C. Musto, and G. Semeraro (Eds.).

[14] A. Pentel. 2015. Employing Think-Aloud Protocol to Connect User Emotions and Mouse Movements. In *6th International Conference on Information, Intelligence, Systems and Applications (IISA2015)*. IEEE Digital Library, https://doi.org/10.1109/IISA.2015.7387970.

[15] A. Pentel. 2017. Emotions and User Interactions with Keyboard and Mouse. (2017). Unpublished.

[16] A. Pentel. 2017. Predicting User Age by Keystroke Dynamics. (2017). Unpublished.

[17] T. A. Salthouse. 1984. Effects of age and skill in typing. *Journal of Experimental Psychology: General* 113, 3 (1984).

[18] J. Vacca. 2007. *Biometric technologies and verification systems.* Butterworth-Heinemann.
[19] J. R. Young and R. W. Hammon. 1989. Method and apparatus for verifying an individual's identity. *US Patent Office* 4805222 (1989).

View publication stats