Biometric User Model for Recognition on the Web

Peter Krátky[*]
Institute of Informatics, Information Systems and Software Engineering
Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava
Ilkovičova 3, 842 16 Bratislava, Slovakia
peter.kratky@stuba.sk

Abstract

Nowadays, the Internet serves to huge masses of users who often browse anonymously or sometimes use duplicate identities. Methods aimed at recognition of users work with machines, but fail to distinguish real persons behind the computers. We present a concept of biometric component as a part of user model for recognition on the web including a framework for assessing quality of features based on input devices. Mouse-based features are compared from different aspects in a case study from e-shop environment. Further, the we propose a method for recognition individuals based on comparing features distributions, which is evaluated from different points of view. Method application as deduplication of visitors records on the web based on mouse usage is also described in the paper. In addition to this, a method for estimation of user characteristics based on input devices usage is proposed. Its evaluation shows modeling of age and gender of users in e-shop environment from mouse usage data.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*input devices and strategies, interaction styles*; D.4.6 [**Management of Computing and Information Systems**]: Security and Protection—*authentication*

Keywords

Recognition of web users, person identification, behavioural biometrics, biometric traits, computer mouse usage biometrics, soft biometric traits, implicit user modeling

1. Introduction

The Web in its beginnings was based on the trust that users contribute using their own identity. But nowadays, anonymous world of the Internet makes it hardly possible to distinguish the truth. Mechanism which assign identities to users are easily avoided and users can hide behind multiple identities. This results in various negative effects, such as trolling or even abuse in the first place. Secondly, duplicated identities degrade services delivered to users. Visitors who change identities cannot have full-featured personalization. Also, website owners have difficulties to improve the user experience with low-quality data. There are mechanisms which are able to identify browsers and machines (e.g. cookies [15], browser fingerprints [7]), but all of them have one common shortage - inability to distinguish physical persons.

In our dissertation thesis, we introduce a novel approach to distinguish and categorize web visitors based on low-level behavioural patterns. Usage of input devices, such as keyboard, computer mouse or phone touchscreen differs among their users. And this has become a part of biometrics, known as behavioural biometrics, which studies behavioural differences of humans [14].

1.1 Input Devices Usage Biometrics

Research in the field of biometrics based on input devices usage is aimed mainly at improving security of information systems. There are two main categories of articles in this field according to which problem it addresses:

- *Static authentication.* The aim is to secure verification of user's identity at login time. The works are focused on typing dynamics of a short predefined or known text to user (e.g. password) [21, 16, 5], moving with the mouse in a predefined manner [9, 2, 20] or unlocking the phone [6].
- *Dynamic authentication.* Identity is verified continuously while the user works with the system. These works study typing an arbitrary text [11] or free, arbitrary movement with the mouse [18, 22, 1].

The performance of such systems is quantified by false acceptance rate of impostors and false rejection rate of legitimate users. Works focused on authentication based on keystroke dynamics show better results than mouse-based methods, often error rates are lower than 5%. Also, few mouse-based and touchscreen-based methods obtained such results. However, comparison using these metrics is

[*] Recommended by thesis supervisor: Assoc. Prof. Daniela Chudá
Defended at Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava on May 4, 2017.

Krátky, P. Biometric User Model for Recognition of Web Users. Information Sciences and Technologies Bulletin of the ACM Slovakia, Vol. 9, No. 2 (2017) 19-24

not very relevant due to different experiment conditions or methodology used.

In our work, we devote to finding duplicate identities and this task is related to the *identification* problem from biometrics, which in context of input devices usage has been explored just marginally. Our purpose is not connecting users with their real world identities, therefore we use more neutral term *recognition* in order to avoid this rather negative meaning.

Among tens of related papers, only a single is devoted to identification [17], based on keyboard dynamics. Methods from the continuous authentication category became a cornerstone for our research, especially those focused on mouse usage. This is because of free and navigational character of browsing the web, when a pointing device is essential. The works differs a lot in various stages of the process of biometric verification. We focused on the preprocessing stage, studied the approaches across the works and we gathered the ideas in one place. For example, we set up a collection of mouse-related features along with their formulas.

Apart from biometric traits with potential to distinguish individuals, there are also *soft biometric traits*, which are low in distinctiveness, but characterizes some group traits [13]. There are a few papers [8, 12, 10], in which automatic way of extracting these traits is presented. The source of data for these works are keyboard dynamics and gender, age group or handedness was classified as well as whether one or two hands were used for typing.

1.2 Goals of the Thesis

An extensive research of papers in the area of input devices biometrics lead us to identification of two open problems. These papers are focused mostly on verification of identity, but assigning identity to one out of many subjects has been studied just marginally, although this is a common task in biometrics field. Another task which deserves attention is extracting so called soft traits from biometrics data, such as age, gender and so on. And both tasks have a potential to be used in the web science. We set up two hypothesis for our research:

- Recognition of users within a website using input device biometrics is a possible task taking into consideration non-trivial number of users,
- Inferring additional information about user from input device usage is also a possible task.

Based on these hypothesis, we identified three goals for our dissertation thesis.

- *Design and evaluate a method for recognition of users on the web according to their work with input devices.* This is the main goal of our thesis. We assume that such recognition on the web is possible with sufficient performance to improve current techniques. Important aspects for the method evaluations are recognition success rate as well time to achieve recognition results. Further, the method should be scalable as well as independent from system domain where it operates.
- *Define properties of the biometric characteristics based on input devices.* Regarding the recognition method, our partial goal is to create a framework for assessing quality of biometric features in order to select suitable ones for verification and recognition in particular domains. The purpose is to make the input devices features comparable from various points of view, such as changes of mouse usage over time or due to hardware replacement. Such framework could help the researchers to improve the stages prior to recognition.
- *Design and evaluate a method for categorizing users.* The third goal is to design and evaluate a method for categorizing users according to input device usage. We hypothesize that there are properties of users which influence their work with the devices and therefore could be estimated, so it could help to overcome the cold-start problem of personalization, for example.

2. Biometric Component of a User Model

The process of building a biometric profile of a user has a lot of common with user modeling process in adaptive systems. Specifically, collecting relevant actions from users in implicit way and inferring conclusions about the user [3]. A suitable abstraction is layered model [4], in which fine-grained raw events are processed into higher-level structures.

Recognition in its essentials requires some unique characteristics which characterize each individual. In our proposed solution, biometric traits become a part of user model and serve as an additional identifier. We denote this part as *biometric component.*

2.1 Proposed Biometric Component Based on Input Devices Data

From the top level of view, process of building the biometric profile has several stages. The input data for the process come from user interface, where computer mouse, touchscreen or another device *events are logged* (e.g. time and code of a pressed key or time and position of the cursor movement). It is worth mentioning that data are not captured onetime, but rather flow in time. The process then continues with *actions composition* in which events are grouped into actions (e.g. movement stroke). This is closely associated with *features extractions* stage when various characteristics are calculated for each action and stored (e.g. movement stroke with velocity, curvature, etc.). Finally, the *biometrics profile construction* takes place and the biometric component is updated in the database.

There are three types of data objects (events, actions, biometric profile) each at the different level of granularity. Higher-level objects require multiple lower-level objects. Therefore data with low granularity are cumulated, stored and pushed forward in batches. These three types of objects correspond to three vertically stacked layers of the biometric component (see Figure 1). As the component stores measured facts about the user, it belongs to evidence layer of the user model.

The upper layer of *biometric profile* holds aggregated data used for the subsequent recognition. We propose representing the profile in form of histograms, where each fea-

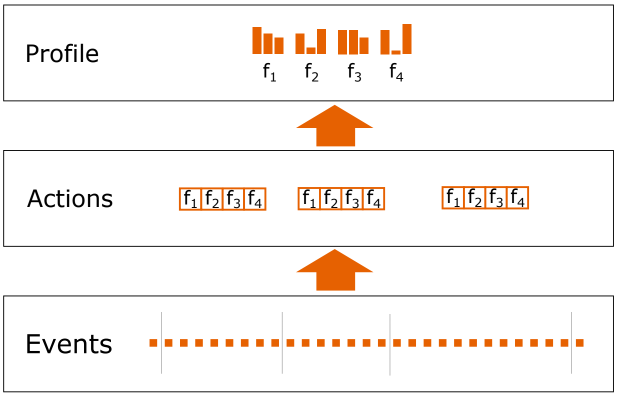


Figure 1: Three vertically stacked layers form the biometric component - fine-grained events at the bottom, higher-level actions in the second layer, biometric profile at the top.

ture has a separate histogram built from values of actions. The reason is to depict distributions of features which are according to our observations characteristic for each user.

2.2 Framework for Assessing Biometric Features of Input Devices Usage

Biometric features have different qualities and this should be taken into account when deciding which characteristics to store in the biometric component. The adjective *biometric* necessarily means that the feature satisfies four requirements: universality (each person should have the trait), distinctiveness (any two persons should differ in trait), permanence (invariant over time), collectability (quantitatively measured) [19].

In the context of pointing devices, the decent number of features could be calculated for everyone using the device, so these features easily fulfil the requirement for universality and collectability. Regarding the other two requirements, we propose description of quality of the feature bf_i as a 4-tuple, formally:

$$Q(bf_i) = (d_i, c_i, p_i, h_i), \tag{1}$$

where d_i stands for distinctiveness of the feature f_i, c_i for consistency (short-term persistence), p_i for long-term persistence and h_i for hardware independence. The last three are components of the permanence requirement.

- *Distinctiveness.* It could be quantified as a probability that two persons differ in the feature. Estimation equals to number of couples with significantly different values (by statistical test) divided by all couples.
- *Consistency.* Estimation equals to number of users with not significantly different values of two samples taken without time delay divided by number of users.
- *Long-term permanence.* It could be estimated as number of users with not significantly different values of two samples taken with time delay divided by number of users.

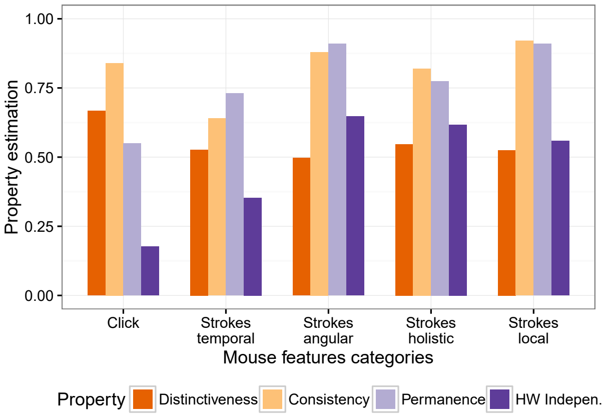


Figure 2: Estimation of mouse-based biometric features properties by categories (median for each category).

- *Hardware independence.* Estimation equal to number of users with not significantly different values of two samples taken with different device divided by number of users.

An desired case is to select the features with the highest sum of the four values. However, biometric features from input device usage are not ideal, so there is usually a trade-off. Features with higher permanence might not have a great value of distinctiveness or short-term consistency and vice versa. Also, estimations of the properties might differ in various domains.

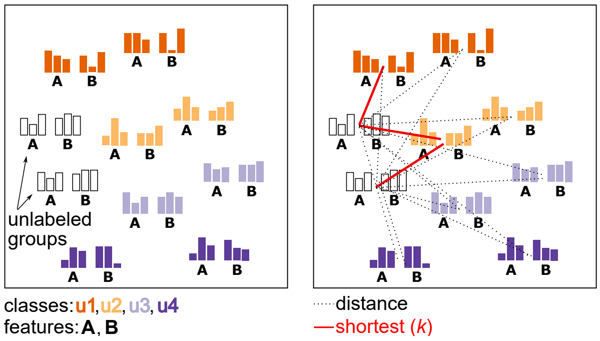
2.3 Mouse-Based Features under Review

Navigational character of the web and privacy issues led us to study pointing devices as a source of data. computer mouse provide us four events: mouse *movement*, mouse *button up* and *button down* and mouse *wheel scrolling*. Events are grouped into actions of suitable granularity, specifically *mouse strokes* (a series of movement events ended with click or pause). Features describing strokes could be calculated from event information, which contains time in milliseconds and coordinates (X and Y).

According to literature we divided all features (as well as few added) into 7 categories: *click* features, *temporal*, *angular*, *holistic*, *local stroke* features, *silence* and *scrolling* features.

To examine the features we conducted a controlled experiment with 25 participants browsing an e-shop while doing several tasks to simulate the shopping activity. It was designed to ensure the same environment for all participants - we provided unified user interface (size, positions of elements) and hardware (monitors, computer mouses). Some of them (19) provided data with hardware replacement, and some (12) did similar activity after one month.

Then we estimated quality properties of the features and compared particular categories. The most distinctive are click features which seem to be consistent in short-term, but rather changing progressively with time. One more advantage of click features is that it requires logging and storing minimum number of events, in contrast to movement features calculated from large number of events. As



(a) (b)

Figure 3: Illustration of the proposed classification method histogram-based kNN - (a) two-dimensional space (features A and B) containing groups of 4 classes and one unlabeled, (b) searching k shortest distances ($k = 3$)

for movement actions, angular features provide high quality in all aspects, similarly local ones do. Scrolling and inactivity seem to be useless for recognition.

We studied also influence of preprocessing on the features. An interesting finding is that the smaller parts of mouse strokes does not have as good properties as full strokes so higher level of grouping is desired. Mouse strokes seem to provide reasonable level detail. Further, we found that approximation of the path has a positive impact on movement features when hardware is replaced.

3. Recognition Using Biometric Component

The biometric component is a kind of identifier that is not completely unique but provides sufficient distinctiveness to differentiate users. The recognition process takes place right after the biometric component is built. In its essentials, the component is matched to the template components stored in the database in order to find the matching one and consequently to determine identity. After the decision is made, an action could be made, for example, merging duplicated identities or splitting records from shared computer.

3.1 Histogram-Based k-Nearest Neighbors

For matching biometric profiles, we propose a modification of k-Nearest Neighbors classifier. It is based on the idea that the particular movement instances (actions described with features) alone do not hold enough information about the user identity, but statistics about group of the instances do. In the database there are multiple groups of instances (one or more for each user) with computed histograms for each feature within the group, denoted as biometric profile above.

In the classification phase, an unlabeled group is compared to all labeled groups and the distance is estimated using a suitable distance metric. We propose using statistics adopted from non-parametric tests to express distance between histograms as most of the features follow non-normal distributions.

Finally, when the unlabeled group is classified, label (identity) of k nearest groups using majority voting scheme is returned as a result.

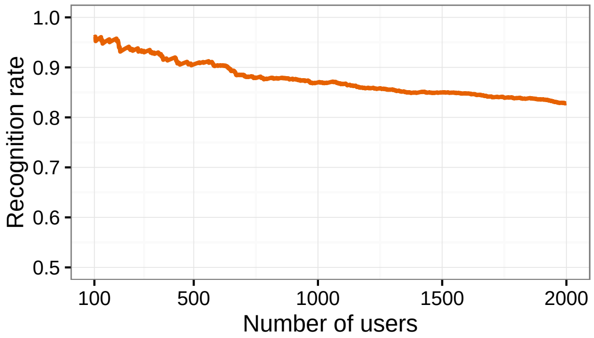


Figure 4: Influence of users pool size on performance (10 features and Kolmogorov-Smirnov statistics as distance metric).

3.2 Method Performance on Mouse-Based Data

We examined several aspects of the proposed method: influence of data availability, level of grouping instances, number of users in the database, applicability in various domains.

We prepared several datasets. The basic dataset was acquired in a tourist portal containing data from 100 active users (providing sufficient amount of data). The extended dataset was collected in the same domain containing more than 2000 active users. The third dataset contains data acquired in three different domains (e-learning system, e=shop and again tourist portal) each containing tens of active users.

Comparing the proposed method with other types of classifiers on the basic dataset shows that histogram-based kNN has a great recognition rate, especially when statistics from Kolmogorov-Smirnov test was used as a distance metric to compare distributions. Recognition rate rises logarithmically with amount of data provided by the user, but 20 movement strokes are required.

As for scalability, the proposed method was able to pick one user out of 2000 with recognition rate over 80%. The rate decreases slowly with increasing number of users. Also, the method was able to operate with similar performance in both e-shop and e-learning domain.

3.3 Deduplication of Website Visits

The proposed recognition method requires static number of classes when determining identity. Some real-world applications might require also making a decision whether the user has its template in database. Such an application is deduplication of website visits, which takes place when users erase identifiers (e.g. cookies) and duplicated identities are created. This affects statistics of the website traffic as returning visitors have status of new ones.

This requires to extend the method to work with dynamic number of classes. To determine whether the user has its template in database or not, we establish threshold t. If the shortest distance returned by the proposed method is greater than the threshold it is considered as a new class (unknown user).

In our work we propose a *dynamic threshold*. When calculating distances, one more distance is calculated between the current and an *average user* (whose biometric profile is based on data from all users) and this distance becomes the new threshold. As distance between two histograms depends also on how smooth histograms are, in this approach the average user serves as an mediator, which eliminates this bias.

In the context of deduplication of website visits, we do not rely solely on mouse biometrics. When comparing biometric profiles, filtering of the users pool is done using device factors, such as screen resolution, operating system, browser, etc. And our mouse-based method is applied to determine whether user has its template from the reduced pool.

For evaluation we used the large dataset with users having variable amount of data (long-tail distribution mostly). We simulated erasing identifier each day during the period of 5 weeks. This method was able to determine number of new and returning visitors more accurately in comparison to common tools based on cookies. Obviously, when number of users rises in the database each day, the performance decrease. We consider using this method for historical data no more than 14 days old.

4. Categorization Using Biometric Component

Another utilization of the biometric component is categorization of users. We propose extending a user model with soft biometrics , which holds high-level information about the user. The soft biometrics are inferred from the biometric component stored in the evidence layer, which is at the lower level. The process of modeling soft biometrics does not use the biometric profile, but the layer of actions below it described with features. To infer the characteristics, machine learning is employed to infer the information from actions characterized with values of features.

Characteristics modeled this way is not based on *what the user browses*, but *how the user browses*. This means that information about context is not necessary and the process of modeling has more general application.

4.1 Soft Biometrics from Mouse-Based Data

In order to explore possibility of modeling soft biometrics we studied differences in features between groups of users. We prepared two datasets. The first contained data from controlled experiment in e-shop from the first part of our study with both males (16) and females (9). The second dataset contained data from 73 students fulfilling a personality questionnaire Big Five that results in quantified five traits.

We sampled several groups of all classes and tested whether differences within and between classes are significant or not using Wilcoxon rank test. As for gender, we found few weak features only, such as deviation of acceleration or average horizontal velocity. Regarding personality, a discriminating feature for people lower and higher in neuroticism was found average curvature change, for instance.

4.2 Estimating Gender and Age in E-Shop

Going further, we performed an experiment in large-scale with goal to estimate gender and age group (two classes). We prepared a dataset containg mouse data browsing a book e-shop from more than thousand users with gender and age label assigned. After selection of active users balancing gender classes, we end up with dataset from 424 visitors.

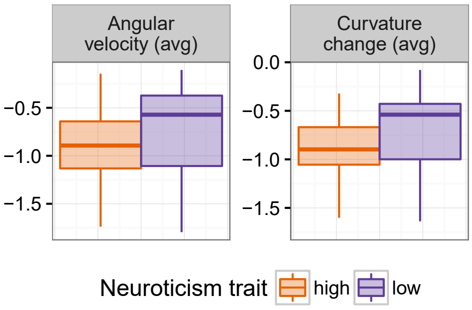


Figure 5: Boxplots illustrating differences of medians between users low and high in the neuroticism trait of the two selected features.

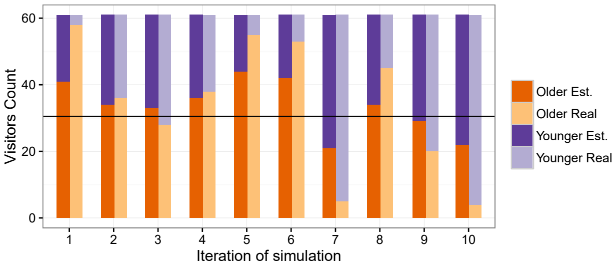


Figure 6: Estimation of major category for age groups. Estimated values are illustrated with darker bars, real values with lighter bars in background.

We trained multiple classifiers and combined them using sum fusion scheme. As for estimating gender we reached F-score 0.6 when using data from 9 visited pages by each user. Age group could be estimated better, with F-score 0.67, based on 10 visited pages.

Although this method does not provide sufficient performance as a standalone solution for modeling, we see a practical application for estimating major category viewing a website. Testing the method to determine whether the major audience was younger or older resulted in 90% success rate.

5. Conclusions

The research in biometrics based on input devices has progressed a lot in identity verification problem. We believe that recognition task also deserves attention, especially in the context of web. Moreover, biometric features might be used for inferring information about users.

Our work presents a novel approach to recognition and categorizations of website visitors based on biometrics. We designed and evaluated the person-level recognition method for the web based on biometric component, specifically input devices usage biometrics. The method was evaluated from different points of view, such as scalability or applicability in other domains. Our next contribution is definition of properties of the biometric features from input devices and a method how to quantify them.

This framework might help researchers to improve performance by enhancing the preprocessing stage. Finally, we proposed extension for modeling user properties with utilization of biometric component.

When fulfilling goals of our dissertation thesis, we have built various datasets from controlled and uncontrolled experiments. This output could be beneficial for the community working in this field of research.

We see another possible direction in users recognition by approaching it as a time series problem. The data could be processed as the flow of angle and velocity changes in time. The combination with the proposed method could improve not only performance but also hardware independence.

Despite the fact that the method was evaluated on the mouse data, we assume that it will work for other types of input devices as well. Especially, mobile devices provide a big opportunity of multiple sources of data. New sources of data and approaches open new research questions not only for recognition issue but also modeling of characteristics of the users.

Acknowledgements. This work was partially supported by the grants No. VEGA 1/0971/11, VEGA 1/0646/15 and APVV-15-0508.

References

[1] A. A. E. Ahmed and I. Traore. A New Biometric Technology Based on Mouse Dynamics. *IEEE Transactions on Dependable and Secure Computing*, 4(3):165–179, 2007.

[2] P. Bours and C. J. Fullu. A login system using mouse dynamics. In *2009 Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pages 1072–1077, Sept 2009.

[3] P. Brusilovsky. Adaptive hypermedia. *User Modeling and User-Adapted Interaction*, 11(1):87–110, 2001.

[4] P. Brusilovsky and E. Millán. User models for adaptive hypermedia and adaptive educational systems. *The Adaptive Web*, pages 3–53, 2007.

[5] D. Chudá and M. Ďurfina. Multifactor authentication based on keystroke dynamics. In *Proceedings of the International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing*, CompSysTech '09, pages 89:1–89:6, New York, NY, USA, 2009. ACM.

[6] A. De Luca, A. Hang, F. Brudy, C. Lindner, and H. Hussmann. 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*, CHI '12, pages 987–996, New York, NY, USA, 2012. ACM.

[7] P. Eckersley. *How Unique Is Your Web Browser?*, pages 1–18. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[8] M. Fairhurst and M. D. Costa-Abreu. Using keystroke dynamics for gender identification in social network environment. In *4th International Conference on Imaging for Crime Detection and Prevention 2011 (ICDP 2011)*, pages 1–6, Nov 2011.

[9] H. Gamboa, A. L. N. Fred, and A. K. Jain. Webbiometrics: User verification via web interaction. In *2007 Biometrics Symposium*, pages 1–6. IEEE, 2007.

[10] R. Giot and C. Rosenberger. A new soft biometric approach for keystroke dynamics based on gender recognition. *International Journal of Information Technology and Management*, 11(1/2):35–49, 2012.

[11] D. Gunetti and C. Picardi. Keystroke analysis of free text. *ACM Trans. Inf. Syst. Secur.*, 8(3):312–347, Aug. 2005.

[12] S. Z. S. Idrus, E. Cherrier, C. Rosenberger, and P. Bours. Soft biometrics for keystroke dynamics: Profiling individuals while typing passwords. *Computers & Security*, 45:147 – 155, 2014.

[13] A. K. Jain, S. C. Dass, and K. Nandakumar. Soft Biometric Traits for Personal Recognition Systems. In *Proc. of International Conference on Biometric Authentication (ICBA)*, pages 731–738, 2004.

[14] A. K. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *IEEE Trans. Cir. and Sys. for Video Technol.*, 14(1):4–20, Jan. 2004.

[15] B. Krishnamurthy and C. E. Wills. Generating a privacy footprint on the internet. In *Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement*, IMC '06, pages 65–70, New York, NY, USA, 2006. ACM.

[16] H.-j. Lee and S. Cho. Retraining a keystroke dynamics-based authenticator with impostor patterns. *Computers & Security*, 26(4):300 – 310, 2007.

[17] F. Monrose and A. Rubin. Authentication via keystroke dynamics. In *Proceedings of the 4th ACM Conference on Computer and Communications Security*, CCS '97, pages 48–56, New York, NY, USA, 1997. ACM.

[18] Y. Nakkabi, I. Traore, and A. A. E. Ahmed. Improving Mouse Dynamics Biometric Performance Using Variance Reduction via Extractors With Separate Features. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 40(6):1345–1353, 2010.

[19] S. Prabhakar, S. Pankanti, and A. K. Jain. Biometric recognition: security and privacy concerns. *IEEE Security Privacy*, 1(2):33–42, Mar 2003.

[20] J. Shelton, J. Adams, D. Leflore, and G. Dozier. Mouse tracking, behavioral biometrics, and gefe. In *2013 Proceedings of IEEE Southeastcon*, pages 1–6, April 2013.

[21] E. Yu and S. Cho. Keystroke dynamics identity verification – its problems and practical solutions. *Computers & Security*, 23(5):428–440, 2004.

[22] N. Zheng, A. Paloski, and H. Wang. An Efficient User Verification System Using Angle-Based Mouse Movement Biometrics. *ACM Transactions on Information and System Security*, 18(3):1–27, 2016.

Selected Papers by the Author

P. Krátky, D. Chudá. Recognition of Web Users with the Aid of Biometric User Model. In *Journal of Intelligent Information Systems – IF 1.0*, submitted.

P. Krátky, D. Chudá. Biometric Properties of Mouse-Based Features on the Web. In *ACM Transactions on Privacy and Security – IF 0.759*, submitted.

P. Krátky, D. Chudá. Fine-tuning Web Traffic Statistics by Deduplication and Splitting of Visitors Records Using Mouse Biometrics In *Proceedings of the 17th International Conference on Computer Systems and Technologies (CompSysTech'16)*, pages 300–306, USA, 2016. ACM.

P. Krátky, D. Chudá. Estimating Gender and Age of Web Page Visitors from the Way They Use Their Mouse In *Proceedings of the 25th International Conference on World Wide Web (WWW '16)*, pages 61–62, USA, 2016. ACM.

D. Chudá, P. Krátky, J. Tvarožek. Mouse Clicks Can Recognize Web Page Visitors! In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*, pages 21–22, USA, 2015. ACM.

D. Chudá, P. Krátky. Mouse Usage Biometrics in eLearning Systems: Detection of Impersonation and User Profiling. In *Journal International Journal of Human Capital and Information Technology Professionals (IJHCITP)*, volume 6, pages 39–50. IGI Publishing Hershey, 2015.

D. Chudá, P. Krátky. Grouping Instances in kNN for Classification Based on Computer Mouse Features. In *Proceedings of the 16th International Conference on Computer Systems and Technologies (CompSysTech '15)*, pages 214–220, USA, 2015. ACM.

D. Chudá, P. Krátky. Usage of computer mouse characteristics for identification in web browsing. In *Proceedings of the 15th International Conference on Computer Systems and Technologies (CompSysTech '14)*, pages 218–225, New York, USA, 2014. ACM.