# How You Type Is Who You Are

Krisztian Buza*, Dora Neubrandt**
* Brain Imaging Center, Research Center for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary
** Budapest University of Technology and Economics, Budapest, Hungary
{buza,dori}@biointelligence.hu

***Abstract*—The increasing interest in person identification based on typing patterns may be attributed to several factors. First, cheap and widely applicable person identification is essential due to wide-spread usage of internet based services, such as online courses or internet banking. Furthermore, introduction of new approaches is necessary because of the continuous development of attack techniques against existing identification methods. The dynamics of typing is characteristic to particular users, while a user is hardly able to mimic the typing dynamics of other users. According to recent observations, person identification based on machine learning using data about the dynamics of typing works surprisingly well. Hubness-aware regression techniques have been introduced recently, however they have not been applied to person identification previously. In this paper, we propose to use EC*k*NN, a hubness-aware regression technique together with dynamic time warping for person identification. We collected time-series data describing the dynamics of typing and used it to evaluate our approach. As baseline we used state-of-the-art time-series classifiers. Experimental results show that the proposed technique outperforms the baselines. In order to assist reproducibility of our work, we publish the data we collected.**

## I. INTRODUCTION

Person identification is an essential task in various applications such as exams of online courses, internet banking, etc. Usual techniques for person identification range from passwords to biometric identification, such as fingerprints, iris-patterns, electroencephalograph-based and electrocardiograph-based person identification [1] [2]. The wide-spread use of the aforementioned online services requires cheap, widely accessible and reliable person identification techniques. Additionally, introduction of new approaches is necessary because of the continuous development of attack techniques against existing identification methods.

The dynamics of typing is characteristic to particular users, and a user is hardly able to mimic the typing dynamics of other users [3]. This makes person identification based on typing patterns especially appealing in cases when the user is not necessarily interested to cooperate, such as testing the identity of students taking exams. For example, Coursera (see: www.coursera.org), one of the world's largest online course providers, identifies users based on the dynamics of their typing, when these users solve quizzes, tests, exams associated with the online courses. Although Coursera provides online courses to over 17 million users worldwide, note that user identification based on typing patterns could be potentially useful in case of regular on-campus courses and exams as well. Furthermore, in cases when high accuracy is required, such as online bank transactions involving large amounts of money, person identification based on typing patterns could be used in combination with conventional identification techniques, such as passwords and authentication codes sent to the user in SMS or e-mail.

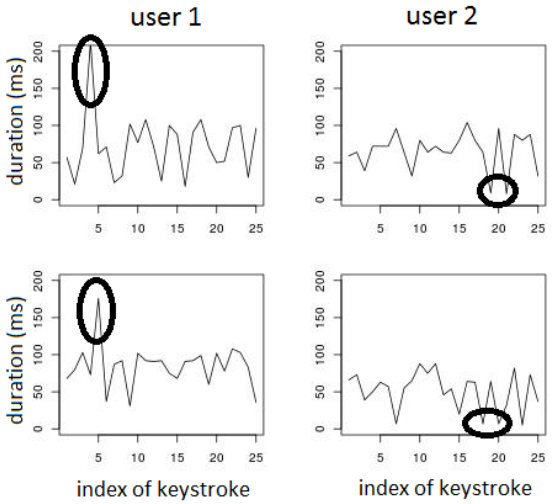


Figure 1. The duration of 25 consecutive keystrokes in case of two different users (left and right) when typing the same text in two different sessions (top and bottom).

Although the dynamics of typing, e.g. the time series of the duration of keystrokes, is characteristic to users, it is obvious that even the same user can not always type with the *exactly* same dynamics. This is illustrated in Fig.1. The figure shows the durations of the first 25 keystrokes in case of typing the same text by two different users. Each of the users typed the same text two times. The time series of user1 are shown in the left of the figure, while the time series of user2 are shown in the right of the figure. As one can see, the time series of the same user are more similar to each other than the time series of different users. In particular, a peak (i.e., an exceptionally long keystroke) close to the 5$^{th}$ position is characteristic to user1, whereas exceptionally short keystrokes close to position 20 are characteristic to user2. In case if we consider a large set of users, such as millions of Coursera-students, it may be difficult and time-consuming for human experts to identify patterns that are able to reliably distinguishing users from each other. Therefore, approaches based on machine learning are required for user identification.

We consider the task of person identification based on the dynamics of typing as a time-series classification problem, for which various approaches have been introduced ranging from neural networks [4] [5] over

Hidden Markov Models [6] to support vector machines [7] and Bayesian networks [8]. However, the 1-nearest neighbour (1NN) classifier with dynamic time warping (DTW) as distance measure was shown to be an extremely competitive classifier, outperforming many complex models, such as neural networks, Hidden Markov Models or "super-kernel fusion scheme" [9] [10]. While the empirical evidence is also justified by theoretical results [11] [12], one of the recently observed shortcomings of nearest neighbour models is their suboptimal performance in the presence of bad hubs [13] [14]. Instance $x$ is called a *bad hub*, if $x$ appears as one of the $k$-nearest neighbours of surprisingly many other instances, but $x$ belongs to a class which is different from the class of those instances that have $x$ as their nearest neighbour. With *hubness*, we refer to the presence of bad hubs, a phenomenon that has been observed in various datasets, including time series datasets [15]. For a more formal definition of bad hubs, we refer to [15], in which hubness-aware classifiers are surveyed and applied to the classification of time series. Because bad hubs are responsible for surprisingly large fraction of the total classification error of nearest neighbour classifiers, as shown in [15], reduction of the detrimental effect of bad hubs can substantially improve the accuracy of time-series classification. Despite the fact that the dynamics of typing can be described by time series and hubness-aware models are among the most promising recent machine learning techniques for time-series classification, they have not been applied to person identification based on typing patterns.

In this paper we propose to use hubness-aware models for the task of person identification based on time series describing the dynamics of typing. We performed experiments on real-world data and show that hubness-aware models outperform prominent time-series classifiers. In order to be able to evaluate our approach, we collected data over several months from different users. In order to assist reproducibility and to motivate further research, we made our data publicly available at http://biointelligence.hu/typing.html and opened a challenge at http://www.biointelligence.hu/typing-challenge/ .

The rest of the paper is organized as follows: Section II describes the details of the proposed approach, while Section III presents our experimental results. Finally, we conclude in Section IV.

## II. Hubness-aware Regression for Person Identification based on Typing Patterns

We base our solution on the wide-spread "classification-via-regression" approach. In particular, we use a regression technique for the person identification task in the following way: for each pair of users ($u$,$v$) we train a separate model. While doing that, we associate time series describing the typing dynamics of user $u$ with label "0". Similarly, the time series of user $v$ are associated with label "1". When a new time series is presented to the trained model, the model outputs a continuous value between 0 and 1 (bounds are inclusive). Values close to 0 (or 1, resp.) indicate that, according to the model, the new time series is more likely to represent the typing dynamics of user $u$ (or $v$, respectively).

Usage of pairwise models, i.e., models that distinguish between two users, is consistent with the circumstances under which person identification problems arise in real-world applications: both in case of the aforementioned online exams and online banking scenario, the user claims an identity and the task is to decide if the claimed identity matches the user's true identity. For example, if the claimed identity is $u^*$, and there are other users in the system, denoted by $u_i$, $1 \le i \le n$, we decide for each $u_i$ whether the typing pattern is more consistent with the typing patterns of $u^*$ than $u_i$. These $n$ decisions can be implemented in parallel if the number of users is high and several computational units are available.[1]

At each of the above decisions, a simple decision threshold of 0.5 could be applied, i.e., given a regression model trained to distinguish between the typing patterns of users $u$ and $v$, if the model outputs less than 0.5 when a new time series is presented to the model, then the decision is $u$, otherwise the decision is $v$. However, the simple threshold of 0.5 may be suboptimal, therefore, we learn the threshold in the following way: once the model is trained, we present the time series of the training set to the model and obtain the output of the model for the training time series. Then, we determine the threshold that gives the highest accuracy on the training data.

Next, we specify the regression model used for the aforementioned pairwise decisions.

### A. Nearest neighbour regression with error correction

We propose to use the $k$-nearest neighbour regression with error correction (EC$k$NN) over the time series representing the dynamics of typing as pairwise decision models. EC$k$NN is a hubness-aware extension of the $k$NN regression. By design it is suitable to various types of data, e.g. vector data, time series, etc., given that an appropriate distance measure between the instances of the dataset is available. As we work with time-series data describing the dynamics of typing, the instances are the time series in our case. Next, we describe EC$k$NN in more detail.

In its training phase, EC$k$NN implements error correction on the training data. In particular, the corrected label $y_c(x)$ of an instance $x$ is defined as

$$y_c(x) = \begin{cases} \frac{1}{|\mathcal{I}_x|} \sum_{x_i \in \mathcal{I}_x} y(x_i) & \text{if } |\mathcal{I}_x| \ge 1 \\ y(x), & \text{otherwise} \end{cases}, \qquad (1)$$

where $\mathcal{I}_x$ denotes the set of training instances that have $x$ as one of their $k$-nearest neighbours and $y(x)$ is the original (i.e., uncorrected) label of instance $x$. When EC$k$NN is applied to predict labels for new instances, it performs $k$-nearest neighbour regression using the corrected labels. That is: for a new instance $x^*$, EC$k$NN searches for the $k$-nearest neighbours of $x^*$ among the training instances and outputs the average of the corrected labels of the neighbours as the estimated label of $x^*$.

Next, we illustrate the error correction mechanism performed by EC$k$NN on a simple example shown in Fig.2. In the figure, training instances are denoted by circles. They are identified by the symbols $x_1 \ldots x_7$. The numeric value (0 or 1) next to each instance shows its

[1] As there might be users with similar typing dynamics, depending on the costs of different types of errors, in order to successfully authenticate user $u^*$ we might allow a few of these pairwise decisions to "fail" in the sense that the model outputs that the typing pattern is more consistent with the patterns of $u_i$ than $u^*$.

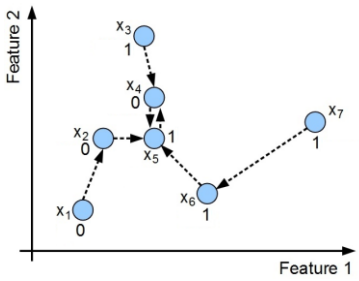


Figure 2. Example used to illustrate error correction.

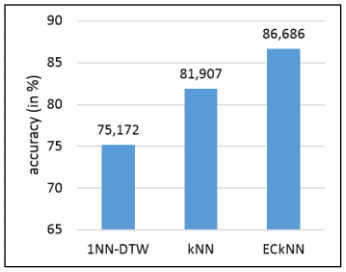


Figure 3. Accuracy of our approach (EC*k*NN) and the baselines.

label. In order to keep the example simple, we use $k = 1$ nearest neighbour to calculate the corrected labels of training instances. In the figure, directed edges point from each instance to its first nearest neighbour. We only present the calculations for $x_4$ and $x_5$ as the procedure is similar in case of the other instances as well. Concretely, the corrected labels of $x_4$ and $x_5$ are:

$y_c(x_4) = (1 / 2) * (1 + 1) = 1$, and

$y_c(x_5) = (1 / 3) * ( 0 + 0 + 1 ) = 0.33$.

For more details about EC*k*NN we refer to [16].

### B. *Dynamic Time Warping*

As we mentioned previously, the dynamics of typing is captured by time series data. In order to use EC*k*NN with time series data, we need to be able to determine the nearest neighbours of time series.

Dynamic time warping (DTW) is a time series distance measure that is robust to elongations and noise. DTW was originally introduced by Sakoe and Chiba [17]. In the last decades, DTW emerged as one of the most prominent techniques in machine learning with time series. Therefore, we propose to use DTW as a distance measure for person identification based on time series describing the dynamics of typing. For a detailed description of DTW, we refer to [15].

## III. Experimental Evaluation

### A. *Data Collection*

We collected time series data describing the dynamics of typing, or *typing patterns* for short, from four different users over several months. Each of the users donated approximately 50 typing patterns, resulting in a collection of 200 typing patterns in total. In each typing session, the users were asked to type the following short text based on the English Wikipedia page about Neil Armstrong:

> *That's one small step for a man, one giant leap for mankind. Armstrong prepared his famous epigram on his own. In a post-flight press conference, he said that he decided on the words just prior to leaving the lunar module.*

In each typing session, we measured both (i) the time between consecutive keystrokes and (ii) the duration of each keystroke, i.e., the time between pressing and releasing each key. We used a self-made JavaScript application and a PHP script to capture the aforementioned time series and to save the data. Note that due to typing errors, the length of typing patterns varies slightly from session to session. In order to encourage future research, we made our data publicly available at http://biointelligence.hu/typing.html .

We note that our protocol, according to which users type the same text in each typing session, corresponds to Coursera's user authentication based on typing dynamics: before writing authenticated tests, Coursera students are asked to type a given short text.

### B. *Experimental protocol*

In order to simulate the scenario in which users provide few typing patterns when they register into a system, we used the first five typing patterns per user as training data. The remaining typing pattern were used as test data in order to evaluate the system.

For the reasons mentioned in Section II, we trained models to distinguish two users, i.e., we trained a model for each pair of users. For each model, we measured the accuracy, i.e., the ratio of correctly classified instances. We report accuracies averaged over all the pairs of users.

We measured the accuracy of our approach, EC*k*NN, and the baselines in case of (i) using only the time series of the times between consecutive keystrokes, (ii) using only the time series of the duration of keystrokes, and (iii) using both of the aforementioned two time series. In the latter case, we combined the output of the two models used in the first two cases by averaging their outputs.

We used the public EC*k*NN implementation from the PyHubs library (http://biointelligence.hu/pyhubs). We set $k = 5$ for EC*k*NN which is in accordance with other works on hubness-aware machine learning [16] [18].

As described in Section I, 1NN-DTW was reported as an extremely competitive time series classifier, therefore we used it as one of the baselines. Additionally, we used *k*-nearest neighbour regression (*k*NN) with DTW and $k$=5 and a decision threshold learned on the training data (see Section II for details of the "classification-via-regression" approach and how we learned the decision threshold).

### C. *Experimental Results*

Fig. 3 and Fig. 4 summarize the results of our experiments. Fig. 3 shows classification accuracy of the combined models that use both types of time series. We observed that EC*k*NN outperforms both baselines statistically significantly according to binomial tests [19] at significance level of $p = 0.001$.

In Fig. 4, we examine the performance of both types of time series in more detail. Interestingly, keypress duration time-series seems to be more informative than the times between consecutive keypresses. Most importantly, the combination of both types of information leads to the best performance out of the examined cases.

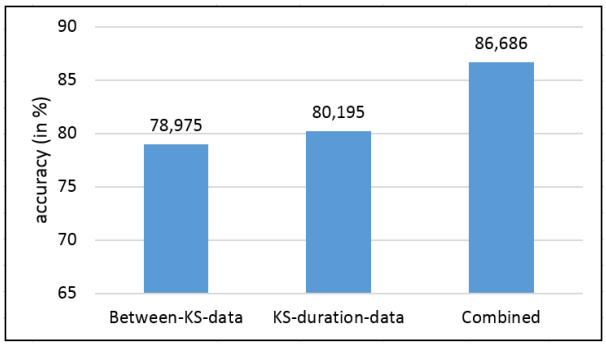


Figure 4. Accuracy of our approach for various types of data.

## IV. Conclusions and Outlook

According to our observations, Georges-Louis Leclerc's famous epigram, *Le style, c'est l'homme*, seems to be valid for typing patterns, in the sense that persons may be identified based on the dynamics of typing. In this paper, we considered the task of person identification based on keystroke dynamics. We proposed to use a hubness-aware regression technique, EC*k*NN, for the person identification task. We compared the results of EC*k*NN with 1NN-DTW and *k*NN regession which are highly competitive baselines. In order to assist reproducibility and to encourage future research, we made our data publicly available and announced a challenge at http://www.biointelligence.hu/typing-challenge/ .

Additionally to EC*k*NN, we tried further hubness-aware regression techniques, namely, EW*k*NN, error-based weighting with *k*NN regression and EWC*k*NN which combines error-based weighting and error correction. As these approaches performed similar to EC*k*NN we omitted to present them for brevity.

W.r.t. the interpretation of the results, we note that we recorded typing patterns over a relatively long period of several months. As we used the first five typing patterns as training data and the remaining typing patterns as test data, the high accuracy we achieved indicates that, for the users we examined, the dynamics of typing was relatively stable over time.

As hubness-aware approaches performed well for the person identification based on the dynamics of typing, we envision that similar techniques will be applied to other tasks as well, such as electroencephalograph-based and electrocardiograph-based person identification.

## Acknowledgments

We thank Ladislav Peška for his comments and remarks on the manuscript. This research was performed within the framework of the grant of the Hungarian Scientific Research Fund – OTKA 111710 PD. This paper was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

## References

[1] S. Marcel, J. Millan, "Person Authentication using Brainwaves (EEG) and Maximum a Posteriori Model Adaptation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 743-752, 2007.

[2] F. Gargiulo, A. Fratini, M. Sansone, C. Sansone, "Subject identification via ECG fiducial-based systems: Influence of the type of QT interval correction," *Computer methods and programs in biomedicine*, vol.121, no.3, pp.127-136, 2015.

[3] F. Monrose, A. D. Rubin, "Keystroke dynamics as a biometric for authentication," *Future Generation Computer Systems*, vol.16, no.4, pp. 351-359, 2000.

[4] F.W.M.H. Wong, A.S.M. Supian, A.F. Ismail, "Enhanced user authentication through typing biometrics with artificial neural networks and k-nearest neighbor algorithm," *Thirty-Fifth Asilomar Conference on Signals, Systems and Computers, IEEE*, vol. 2, 2001.

[5] A. Nanopoulos, R. Alcock, Y. Manolopoulos, "Feature-based Classification of Time-series Data," *International Journal of Computer Research*, vol.10, nr.3, pp. 49-61, 2001.

[6] S. Kim, P. Smyth, S. Luther, "Modeling waveform shapes with random effects segmental hidden Markov Models," Technical Report, UCI-ICS 04-05, 2004.

[7] D. Eads, D. Hill, S. Davis, S. Perkins, J. Ma, R. Porter, J. Theiler, "Genetic Algorithms and Support Vector Machines for Time Series Classification," *Society of Photo-Optical Instrumentation Engineers* (SPIE) Conference Series, vol. 4787, pp. 74–85, 2002

[8] K. Buza, L. Schmidt-Thieme, "Motif-based classification of time series with Bayesian networks and SVMs," in *Advances in Data Analysis, Data Handling and Business Intelligence*, pp. 105-114, 2010.

[9] X. Xi, E. Keogh, C. Shelton, L. Wei, C.A. Ratanamahatana, "Fast time series classification using numerosity reduction," *Proceedings of the 23rd International Conference on Machine Learning*, ACM, pp. 1033-1040, 2006.

[10] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, E. Keogh, "Querying and mining of time series data: experimental comparison of representations and distance measures," *Proceedings of the VLDB Endowment*, vol.1 no.2, pp. 1542-1552, 2008.

[11] G. H. Chen, S. Nikolov, and D. Shah, "A latent source model for nonparametric time series classification," in *Advances in Neural Information Processing Systems*, 2013, pp. 1088–1096.

[12] L. Devroye, L. Györfi, and G. Lugosi, *A probabilistic theory of pattern recognition*. Springer Science & Business Media, 2013, vol. 31.

[13] M. Radovanovic, A. Nanopoulos, and M. Ivanovic, "Hubs in space: Popular nearest neighbors in high-dimensional data," The Journal of Machine Learning Research, vol. 11, pp. 2487–2531, 2010.

[14] N. Tomasev, M. Radovanovic, D. Mladenic, and M. Ivanovic, "A probabilistic approach to nearest-neighbor classification: Naive hubness bayesian knn," in Proc. CIKM, 2011.

[15] N. Tomasev, K. Buza, K. Marussy, and P. B. Kis, "Hubness-aware classification, instance selection and feature construction: Survey and extensions to time-series," in *Feature selection for data and pattern recognition*. Springer-Verlag, 2015.

[16] K. Buza, A. Nanopoulos, and G. Nagy, "Nearest neighbor regression in the presence of bad hubs," *Knowledge-Based Systems*, vol. 86, pp. 250–260, 2015.

[17] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[18] N. Tomasev and D. Mladenic, "Nearest neighbor voting in high dimensional data: Learning from past occurrences," *Computer Science and Information Systems*, vol. 9, pp. 691–712, 2012.

[19] S.L. Salzberg, "On compating classifiers: Pitfalls to avoid and a recommended approach," *Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 317–328, 1997.