# A new Brazilian hand-based behavioural biometrics database: Data collection and analysis

DA SILVA, V.R., DE ARAÚJOSILVA, J.C.G. and DA COSTA ABREU, Marjory <http://orcid.org/0000-0001-7461-7570>

# A new Brazilian hand-based behavioural biometrics database: Data collection and analysis

**Valmiro Ribeiro Da Silva, Julliana C.G. De Araújo Silva, Márjory Da Costa-Abreu***

*DIMAp-UFRN, Natal/RN. Brazil, marjory@dimap.ufrn.br

**Keywords:** Brazilian biometrics database, Keyboard-keystroke Dynamics, Touch-keystroke Dynamics, Online Handwritten Signature

## Abstract

The technologies of biometrics provide a variety of powerful tools to determine or confirm individual identity while, more recently, there has been considerable interest in using soft biometrics (personal information which is characteristic of, but not unique to, individuals) in the identification task. Although the area has seen an increase advance in the last decade, there is still the need to understand the existing demographics differences. More importantly, though, is the acquisition of new and relevant multimodal databases that can be used for new investigations. Therefore, this paper presents a protocol for the collection of the first hand-based multimodal biometrics database with keystroke dynamics, keystroke-touch dynamics and handwritten signature in Brazil.

## 1 Introduction

The use of mobile devices, social networking and electronic messaging are unstoppable phenomena, with many benefits but also dangers in their continuing spread, especially among vulnerable groups. The fast growing and development of such computational systems in recent years have provided their quickly popularisation in several areas of society. When we think about these systems being used in areas such as banking and business, it becomes inevitable not deal with a lot of users and with the inherent necessity of authentication [4].

There are several ways to perform the authentication of users in mobile devices, such as pin access number or the use of biometric data. They vary widely in security, accuracy and acceptability for the population, the pin numbers are easy to be forgotten or stolen, but, that is not the case for the biometric data, for example [15].

The use of biometric-based automatic authentication systems is relatively well accepted [22]. However, one of the main problems in exploring more these type of data is the lack of meaningful and representative databases [21]. For some modalities, such as face [3], iris [12] or fingerprint [13], there are a considerable amount of relevant data. However, when looking for behavioural-based databases, especially multimodal, it can be challenging.

The main motivation behind this work is to collect a new open database of biometrics, where every modality included is hand-based, thus increasing the number of sources of hand biometric data available to research, and providing a database filled with new demographic information. This database will include keyboard-keystroke dynamics, touch-keystroke dynamics and online handwritten signature data.

Physiological characteristics tend to be more robust when continuous authentication is concerned, although some modalities like voice patterns and handwriting generally offer a manageable level of uniformity. On the other hand, having more than only one hand-based behavioural modality can lead us to accomplish new discoveries. It is possible to find in the literature indications that combining keystroke with other kind of biometric data or even with soft biometrics data can make accuracies more reliable and give us new types of information [5, 9].

We believe that by investigating different types of keystroke dynamics (in our case touch and keyboard) as a hand-based biometric as well as online handwritten signature can help us to understand if, collecting more than one kind of biometric (called modalities) that uses the hand to provide information at the same time can influence on the data.

The main goal of this work is to collect a new Brazilian hand-based database with keyboard and touch keystroke dynamics and handwritten signature. We have selected a fixed set of words used in Portuguese and English (for both keystroke datasets), chosen carefully to provide good and reliable information during the collection process. We will also present some initial results analysing the predictability of this new database.

## 2 Hand-based behavioural biometric modalities

Since our goal in this work is to propose a new protocol and collect a new hand-based multimodal database with Brazilian characteristics, it is important to understand what has been done in the literature regarding keyboard-keystroke dynamics, touch-keystroke dynamics and online handwritten signature. This section will present the relevant literature review of the three hand-based modalities that were collected in for this paper.

- Keyboard-keystroke dynamics (also known as keyboard dynamics or only keystroke dynamics) is the study of

the unique timing patterns embedded in an individual's typing and most often developed in a way characteristic of that individual, hence the use of keyboard dynamics as a biometrics-based identification modality. Processing of such data typically includes extracting keystroke timing features such as the duration of a key press and the time elapsed between successive key presses [7, 8].

Two keys typed one after the other are called a digraph, and three consecutively typed keys are called a trigraph [8]. The use of accents, upper and lower case and special characters, as well as the error of the user, inside the collect text must be taken into consideration because allowing those characteristics can make implementation more complicated.

- Touch-screen-keystroke dynamics evaluates the typing capabilities of a user when he/she types directly on a touch-screen mobile device. It can be considered similar to keyboard keystroke dynamics in the sense that it analyses the time between each key as well as the duration with which each key is pressed [5, 2].

  Despite the fact that the features analysed by this modality are similar to the previously cited, we can still find a considerable difference because the measurements taken by the keyboard are mechanical and the ones collected by the touch screen are software-based [20].

- Online handwritten signature can be considered as the basis of one of the oldest identification methods ever used. Signature-based identification/verification processes have played a very important role in document forensic analysis [14], and many studies have shown that for particular applications, it can be a reliable modality [11].

  This specific type of signature is written with an electronically instrumented device and the dynamic information (pen tip location through time) is usually available at high resolution, even when the pen is not in direct contact with the paper [19]. It is an especially user-friendly modality and well accepted socially, because of its familiarity and the fact that it is a non-invasive (and non-intrusive) modality. It has long history of use in forensic environments (most of the applications in this particular domain use off-line processing for obvious reasons) [1].

  It is a modality which offers the possibility of "natural revocability" [23], which in a world worried about privacy and identity protection can be very useful.

There are several databases that collect either of the listed modality and, in very few cases, there are databases with a combination of two of them. However, to our knowledge, this is the very first database that has combined the three modalities from the same users.

## 3 Data collection protocol

Our data collection protocol is composed by a set of tasks, starting with a questionnaire to be applied to every user, in order gather demographic information, such as gender and age. As a matter of convenience, the first modality collected was the keyboard-keystroke, because the users needed to user a desktop computer. The second modality was the keystroke-touch-screen modality and the last was the online handwritten signature, both collected using a tablet.

Initially, the user was asked to answer some demographic related questions, such as which age band they were: $< 25$, $25-40$, $40-60$ or $> 60$; what is their gender: Male or Female; their handedness: left or right handed, or ambidextrous and their level of familiarity with a keyboard: little, average or high. In order to measure the hand size of every user, we have used a hand-size model in the same scale for everyone.

The protocol for the keyboard-keystroke and the keystroke-touch-screen were very similar. We have used fixed text, because it allowed us to make every user give the same kind of data, helping with comparison analysis to be perform later. For both modalities, we have collected the dwell time and flight time of each word described in the protocol, but we only ignored errors for the keyboard-keystroke dynamics modality.

After the questionnaire, each user started with our main task related with keyboard keystroke dynamics. We have selected 20 relevant words from the most frequently used words in the Brazilian Portuguese as well as a set of numbers. Some of the selected words are written in the same way for both Brazilian Portuguese and English which makes our database interesting and comparable with other databases.

We have taken into consideration a few aspects of the languages in order to choose the words collected. The word must have very common digraphs from the Brazilian Portuguese, and, if possible, a common digraph from English. All words from this protocol were carefully analysed to provide good information significance. The list of words used along with its justification can be seen below.

- *america – Chosen because it is an English cognate and it has the digraph 'er', very common in Portuguese, especially with verbs;*
- *internet – Chosen because it is a very common word in both English and Portuguese, having the same meaning, and it has important digraphs like 'in' and 'er';*
- *coisa – Chosen because it is one of the most used words in Portuguese and it has some important digraphs like 'oi', 'is' and 'sa';*
- *normal – Chosen because it is an English cognate and it has some important digraphs like 'no', 'or', 'al';*
- *fazer – Chosen because it is one of the most used verbs in Portuguese, like the English verb "to do";*
- *homem – Chosen because it is one of the most used words in Portuguese;*
- *carro – Chosen because it is a common word in Portuguese and all digraphs are important, mostly 'ar' and and the use of the same letter as a digraph 'rr';*

- *porque – Chosen because it is one of the most used words in Portuguese and it has important digraphs, like 'or' and 'qu';*
- *case – Chosen because it is an English false cognate and it has the important digraph 'as';*
- *video – Chosen because it is a very common in both English and Portuguese, having some important digraphs like 'de' and 'eo';*
- *jesus – Chosen because it is a first name in both languages, having 'us' as an important digraph;*
- *mouse – Chosen because it is a word that exists in both languages and it has important digraphs like 'ou', 'us' and 'se';*
- *felicidade – Chosen because it has important digraphs like 'el', 'li', 'ci', 'id', 'da', 'de';*
- *pequeno – Chosen because it is a common word in Portuguese and it has important digraphs, like 'no' and 'qu';*
- *primeiro – Chosen because is a long word in Portuguese and it has important digraphs, such as 'ro', 'pr' and 'me';*
- *zoom – Chosen because it is a common word in both languages and it has important digraphs, such as 'om' and the use of the same letter as a digraph 'oo';*
- *selfie – Chosen because it is a common word in both languages and it has important digraphs, like 'se';*
- *ultimo – Chosen because it is one of the most used words in Portuguese and it has important digraphs, such as 'mo' and 'ti';*
- *mulher – Chosen because it is one of the most used words in Portuguese;*
- *cuba – Chosen because it is a common word in both languages;*

After the collection of the described words, we have also collected the set of numbers "0168245739". It is important to remember that we do not use capital letters, special characters or accent within or set of words, because this kind of feature may interfere with our analysis, increasing the processing phase, and making the data more complex. Also, we have used an universal QWERTY keyboard to collect the data. An universal QWERTY keyboard can be easily found, and the decisions to work without considering special characters, accents and capital letters make the process of replicate the experiments more simple and possible of replication.

After the keyboard-keystroke dynamics collected is finished, the user types exactly the same words in the tablet and after that, the user provides three samples of its own signature. The signature was written on a given area where the writer uses the tablet pen. The information is only recorded when the pen is in contact with the screen of the tablet, resulting in a file that contains the collected data. We have collected data from 120 people. Each volunteer signed their full names three times in order to minimise the natural variation that can occur every time we sign. If there was any problems with the signature, the user was asked to sign again.

Every software used in the process was tested more than once to verify that everything was working properly, and after the collecting process started, the time spent testing and rethinking our methods has been proved very valuable.

## 4 Data analysis

In order to produce a meaningful database for both keystroke databases, we had to perform a deep analysis of all the digraphs that we could use and that were representative enough for our proposed system.

We have also decided to use only a part of the database for these initial experiments, thus, we have selected 77 users from the same collection session and the distribution of data can be seen in Table 1.

Table 1. Demographic information about the users

| **Age** | $< 25$ | $> 25$ |
|---|---|---|
| Qtd | 70 | 7 |
| **Gender** | Female | Male |
| Qtd | 20 | 57 |

After some experiments, we have decided to produce three samples to each user (the same number so we can combine this samples with the three signatures samples collected). We have some digraphs that were the same, but we needed to use some digraphs 'per similarity' which means that they were chosen as similar because one of the keys is situated next to the other in the keyboard used. Our final list of digraphs used can be seen below.

- sample 1 = ME(america), ER(america), RI(america), IC(america), CA(america), IM(primeiro), IR(primeiro), SE(case), MO(mouse), OO(zoom), DE(video), EL(felicidade), RM(normal) and UE(porque)
- sample 2 = ME (primeiro), ER(internet), RI(primeiro), IC(felicidade), CA(case), OM(zoom), OR(normal), SE(mouse), MO(ultimo), RR(carro), DE(felicidade), EL(selfie), EM(homem) and UE(pequeno)
- sample 3 = ME(homem), ER(fazer), RO(primeiro), IC(america), CA(carro), IM(ultimo), OR(porque), SE(selfie), NO(normal), OO(zoom), DE(video), EL(felicidade), EN(pequeno) and UE(porque)

For each keyboard-keystroke database, we have used the dwell time of the first key, the flight time the user took to go from the first key to the second key and the dwell time of the second key.

For each touch-keystroke database, we have used only the flight time the user took to go from the first key to the second key.

For the signature database, we have used the features presented in [16, 17]. The features are listed below.

- SIGDIST - The signature length.
- AVXV and AVYV - Average velocities in the X and Y planes
- VEL1Y and VEL1X - Mean velocities/maximum velocities in the Y and X planes
- VEL2 and VEL3 - Minimum velocities (different from zero) in the X and Y planes/ Average velocities in the X and Y planes
- VEL4X and VEL4Y - First instance of velocities different from 0 in the X and Y planes
- VEL5 and VEL6 - Average velocities in the X and Y planes / maximum velocities in the X and Y planes
- VELCOR - Correlation between velocities in the X and Y plane
- INITDIR - $\frac{X_{max}-X_{min}}{Y_{max}-Y_{min}}$
- XSIZE - $X_{max} - X_{min}$
- YSIZE - $Y_{max} - Y_{min}$
- TOTALTIME - The total time taken to sign
- VELXZERO and VELYZERO - Total number of samples when velocities = 0 in the X and Y planes.
- AVPRESS - The average pressure.
- PIXELCENX and PIXELCENY - The mean X and Y positions of all signature pixels (pixels forming the signature ink).
- PTD - Total distance in mm travelled by the pen in forming the signature.
- HWRATION - Ratio of signature height to width.
- SET - The execution time (in seconds) to draw the signature.
- DCHANGE - The number of times the signature changes direction
- DIST1 - The total length of signature
- DUR1 and DUR3 - The total time the pen stays in one side of CENTCROSS-X and CENTCROSS-Y.
- DUR2 and DUR4 - The total time the pen stays in the other side of CENTCROSS-X and CENTCROSS-Y.
- TIME3 and TIME5 - The time the pen stays at maximum speed in one side of CENTCROSS-X and CENTCROSS-Y.
- TIME4 and TIME6 - The time the pen stays at maximum speed in the other side of CENTCROSS-X and CENTCROSS-Y.

These are considered very representative e common used signature features and we have selected a good balance of static and dynamic features so we can achieve a realist classification result.

## 5 Classification analysis

In order to analyse the predictability of our newly collected database, we have run some classification experiments with different traditional algorithms as well as with different configurations of the data.

We have chosen three well known classifiers from the Weka toolbox [1] that can be described as follows:

- Multi-Layer Perceptron (MLP) [18] which is a Perceptron neural network with multiple layers. The output layer receives stimuli from the intermediate layer and generates a classification output. The intermediate layer extracts the features, their weights being a codification of the features presented in the input samples, and the intermediate layer allows the network to build its own representation of the problem. Here, the MLP is trained using the standard backpropagation algorithm [10] to determine the weight values.
- Support Vector Machines (SVM) which is based on an induction method that minimises the upper limit of the generalisation error related to uniform convergence, dividing the problem space using hyperplanes or surfaces, splitting the training samples into positive and negative groups and selecting the surface which keeps more samples.
- K-Nearest Neighbours (KNN) [6] which works by having the training set seen as a composed of n-dimensional vectors and each element represents a point in n-dimensional space. The classifier estimates the $k$ nearest neighbours in the whole dataset based on an appropriate distance metric (Euclidian distance in the simplest case). The classifier checks the class labels of each selected neighbour and chooses the class that appears most in the label set.

Table 2 presents all the results of the individual modalities, as well as their two-a-two combination and the identification using all three modalities at once. Those results are very interesting. But we need to understand the individual results first in order to be able to analyse the combinations.

The online handwritten signature is a well known modality and their most used features are diverse and representative

[1] www.cs.waikato.ac.nz/ml/weka/

Table 2. All the accuracy % results for the individual modalities as well as their combination

| **Algorithms** | Signature | Touch | Keyboard | Touch + Keyboard | Touch + Signature | Keyboard Signature | All modalities |
|---|---|---|---|---|---|---|---|
| KNN | 82.0175 | 98.045 | 100 | 100 | 84.5629 | 95.6428 | 100 |
| SVM | 90.6678 | 92.2078 | 60.1732 | 100 | 83.632 | 95.7654 | 100 |
| MLP | 96.9298 | 81.8182 | 92.6407 | 98.6842 | 96.0526 | 92.5439 | 98.6842 |

enough for reaching very good results with such a considerable database size. Since we have used a large set of features, it is not surprising that we have managed to archive, with all three classifiers, very good results. This can imply that we do not need a great quantity of features for online handwritten signature classification.

When we analyse the keystroke results (both touch and keyboard), it is possible to note a very different performance even though we are using very similar characteristics as features for the classification. The keyboard tends to present better results for the KNN and the MLP, when compared with the touch, but the SVM behaves better for the touch instead. This can indicate that the neural-based and the linear-based classifiers need more information. But then again, these results will vary when we select different digraphs.

When we analyse the combination of the modalities the conclusions achieved are:

- There is always an improvement when we combine the features of both keystroke modalities when we compare them with the use of the individual modalities.
- When we compare the combination of any of the keystroke modalities and the signature with their use individually, we see that this combination did not improve the performance, in fact, we achieve worse results when comparing any of those results with the individual signature modality.

And lastly, not surprisingly, when we use all the features from all the modalities, we can see that we achieve almost perfect performance. However, those results are likely to hide some problems since we have some discrepancy in the two-a-two combination results.

It is very interesting that we have managed to reach such high accuracies in our initial experiments and this can indicate that this is a very promising database to be explored. We understand these are initial results and that we need to perform a more detailed investigation, but we believe that since this is the very first hand-based behavioural modality to be collected, specially in Brazil, it has the potential of shedding light to new demographic information until recently unknown.

## 6 Final remarks

This presented a new protocol to collect multiple modalities of hand-based biometrics, with keyboard-keystroke dynamics, touch-keystroke dynamics and online handwritten signature.

Over this work many of our initial thoughts were confirmed. The decision to put an questionnaire within the required tasks for the protocol showed good results, but also showed some flaws, most coming from how we sorted the possible answers, impacting in how we visualise the different groups we described by each questionnaire question.

We have run some initial classification experiments, using each dataset on its own and also an analysis of how they perform when used together. We can say that this database has a lot of potential and our future investigations will aim at understanding the best features to each modality that need to be used and, indeed, if we need the combination of these modalities at all.

This kind of analysis can lead us to answer questions like what group of keys containing digits the volunteers used to type the sequence of digits inside the Keystroke Dynamics software. Now we have enough data to start an analysis considering more in depth all the modalities and their relationship with each other.

This expansion of our analysis will cover features like variance and co-variance between samples, cross tables showing different features to discover characteristics in common from different groups. Qualitative and quantitative analysis could show us interesting unknown results extracted from our current data. An analysis crossing our data with other already existing similar data can be also done, showing demographic differences between Brazilian Portuguese writers and other languages writers.

## References


[1] M.C.C. Abreu and M. C. Fairhurst. Improving forgery detection in off-line forensic signature processing. In *The 3rd International Conference on Imaging for Crime Detection and Prevention*, ICDP 2009, 2009.

[2] S. J. Alghamdi and L. A. Elrefaei. Dynamic user verification using touch keystroke based on medians vector proximity. In *The 7th International Conference on Computational Intelligence, Communication Systems and Networks*, CICSyN, pages 121–126, June 2015.

[3] N. Alyuz, B. Gokberk, and L. Akarun. A 3d face recognition system for expression and occlusion invariance. In *The 2nd IEEE International Conference on Biometrics: Theory, Applications and Systems*, BTAS 2008, pages 1–7, September 2008.

[4] J. Angulo and E. Wastlund. Exploring touch-screen biometrics for user identification on smart phones. In J. Camenisch, B. Crispo, S. Fischer-H?bner, R. Leenes,

and G. Russello, editors, *Privacy and Identity Management for Life*, volume 375 of *IFIP Advances in Information and Communication Technology*, pages 130–143. Springer Berlin Heidelberg, 2012.

[5] H. Aronowitz, M. Li, O. Toledo-Ronen, S. Harary, A. Geva, S. Ben-David, A. Rendel, R. Hoory, N. Ratha, S. Pankanti, and S. Nahamoo. Multi-modal biometrics for mobile authentication. In *IEEE International Joint Conference on Biometrics*, IJCB, pages 1–8, September 2014.

[6] A. Arya. An optimal algorithm for approximate nearest neighbors searching fixed dimensions. *Journal of ACM*, 45(6):891–923, 1998.

[7] S.P. Banerjee and D. Woodard. Biometric authentication and identification using keystroke dynamics: A survey. *Journal of Pattern Recognition Research*, 7(1), 2012.

[8] F. Bergadano, D. Gunetti, and C. Picardi. User authentication through keystroke dynamics. *ACM Transactions on Information and System Security (TISSEC)*, 5(4):367–397, 2002.

[9] S. Bhatt and T. Santhanam. Keystroke dynamics for biometric authentication a survey. In *International Conference on Pattern Recognition Informatics and Mobile Engineering*, PRIME, pages 17–23, February 2013.

[10] Y. Chauvin and D.E. Rumelhart. *Backpropagation: theory, architectures, and applications*. L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 1995.

[11] J. Coetzer, B.M. Herbst, and J.A. Du Preez. Off-line signature verification: A comparison between human and machine performance. In Guy Lorette, editor, *The 10th International Workshop on Frontiers in Handwriting Recognition*, IWFHR 2006. Université de Rennes 1, 10 2006.

[12] D.M. Daniel and B. Monica. Person authentication technique using human iris recognition. In *The 9th International Symposium on Electronics and Telecommunications*, ISETC 2010, pages 265–268, November 2010.

[13] S.C. Dass and A.K. Jain. Fingerprint classification using orientation field flow curves. In *Indian Conference on Computer Vision, Graphics and Image Processing*, ICVGIP 2004, pages 650–655, 2004.

[14] M. C. Fairhurst. Document identity, authentication and ownership: The future of biometrics verification. In *The 7th International Conference on Document Analysis and Recognition*, ICDAR 2003, pages 1108–1116, Washington, DC, USA, 2003. IEEE Computer Society.

[15] T. Feng, X. Zhao, N. DeSalvo, Z. Gao, X. Wang, and W. Shi. Security after login: Identity change detection on smartphones using sensor fusion. In *IEEE International Symposium on Technologies for Homeland Security*, HST, pages 1–6, April 2015.

[16] R.M. Guest. The repeatability of signatures. In *The 9th International Workshop on Frontiers in Handwriting Recognition*, IWFHR 2004, pages 492–497, Washington, DC, USA, 2004. IEEE Computer Society.

[17] R.M. Guest. Age dependency in handwritten dynamic signature verification systems. *Pattern Recognition Letter*, 27(10):1098–1104, 2006.

[18] S. Haykin. *Neural networks: a comprehensive foundation*, volume 13. Cambridge University Press, New York, NY, USA, 1999.

[19] A.K. Jain, F. Griess, and S. Connell. On-line signature verification. *Pattern Recognition Letter*, 35(1):2963–2972, 2002.

[20] N. Jeanjaitrong and P. Bhattarakosol. Feasibility study on authentication based keystroke dynamic over touchscreen devices. In *The 13th International Symposium on Communications and Information Technologies*, ISCIT, pages 238–242, September 2013.

[21] J. Kiltter and N. Poh. Multibiometrics for identity authentication: Issues, benefits and challenges. In *The 2nd IEEE International Conference on Biometrics: Theory, Applications and Systems*, BTAS 2008, pages 1–6, September 2008.

[22] H. Lu, F. Clart-Tournier, C. Chatwin, and R. Young. Worldwide authentication and tracking: a secured mobile payment system implemented using a biometrics approach. In *The IASTED International Conference on Communication Systems, Networks, and Applications*, CSNA 2007, pages 130–135, Anaheim, CA, USA, 2007. ACTA Press.

[23] A.B.J. Teoh and C.T. Yuang. Cancelable biometrics realization with multispace random projections. *IEEE Transactions on Systems, Man, and Cybernetics: Part B*, 37(5):1096–1106, 2007.