The influence of dataset quality on the results of behavioral biometric experiments

Pawel Kasprowski

Institute of Informatics
Silesian University of Technology
44-100 Gliwice, Poland
kasprowski@polsl.pl

Ioannis Rigas

Physics Department
University of Patras
26504 Rio, Patras, Greece
rigas@upatras.gr

Abstract: This paper explores some aspects that are involved during the construction of reliable benchmark sample databases for novel behavioral biometric identification methods, such as the data quality, the recording patterns and the post processing procedures that may be applied on the data. A large collection of eye movement samples was employed as a test case. It was recorded under a variety of settings and processed with a number of different approaches. Our analysis reveals that there are specific features during the construction of a database that may significantly influence the final identification performance. It also leads on the establishment of some guidelines, which can be generalized on other behavioral biometric methods, regarding the factors that should be taken into consideration during the creation, the description and the processing of a database of biometric samples.

1 Introduction

Whenever a novel biometric identification method is proposed, two of the most important factors that need to be evaluated are its distinctiveness and repeatability. Distinctiveness refers to the extent of difference in measurements among different people. Repeatability measures the degree of similarity in biometric samples taken from the same person. The latter is especially problematic in the case of biometric identification based on behavioral cues. Human behavior is susceptible to endogenous and exogenous factors that affect its characteristics over time. For example, the way the persons behave is influenced by their mood - people behave differently when they are sad, angry or just tired. Such type of dependencies should be taken into consideration during the experiments for the construction of a biometric database, since they may affect considerably both the performance and the consistency of a behavioral biometric identification system.

One of the main problems encountered in novel biometrics is the absence of reference databases on which the methods can be evaluated, contrary to the case of well-established methods [Ca11][Or03]. Hence, the prospective researchers need to set their own experiments and create their own database of measurements, which is often a difficult and time consuming task. Only after creating the database of measurements, the researchers are able to perform their identification or authorization algorithms and publish their performance rates. As the published results are dependent on the dataset collected, it is very important for the samples database to have been appropriately constructed. Two main distinctive features that need to be considered during the construction of such a database are: the overall number of samples and the number of different individuals that participated in the experiments. A large number of individuals gives the opportunity to properly evaluate the distinctiveness of a method. Similarly, a large number of samples per individual helps in the evaluation of repeatability. In most papers these two factors are used as the most important and - unfortunately in many cases - the only metrics for measuring the dataset quality. As we will show, there are several other aspects worth to be inspected, since their significance to the final results is not negligible.

The quality of measurements is reportedly one of the major problems of biometric research [Wa02][Ga05]. There are general recommendations on how to control factors that influence performace i.e. volunteers selection or test size. It is also well known that noisy, low quality samples influence the overall result [GT07][Ja00]. It is worth mentioning that quality of samples (named also 'corpora quality' [Wa02]) is something different than quality of the database (meta data quality). There are several papers that analyse quality of dataset i.e. registration procedures, storage procedures, removing odd samples procedures [MW02][HK06]. On the contrast, the impact of the quality of measurements for the case of behavioral biometrics is a less researched domain [De95]. As behavioral biometric methods measure person's activity in a given time, the time delay between successive subject measurements also appears as an important property during the dataset construction. The problem is somehow similar to a 'template aging' effect, well known in biometrics but it concerns much shorter periods of time.

In our paper, we emphasize that commonly published properties such as the number of samples and the number of participants are not enough for a complete assessment of the quality of a biometric database. We propose other possible factors that should also be taken into consideration (especially time related metrics) and we demonstrate the impact of these factors on the overall identification results. To the best of our knowledge, there are no generally accepted quidelines of dataset preparation for behavioral biometric identification experiments so this paper aspires to present the influence of certain parameters during the construction of samples datasets and serve as a useful tool during the setup of novel experiments (focused, but not limited to eye movement biometrics) in the future.

2 Eye movements biometrics

The idea of using eye movements for human identification is almost ten years old, with several publications showing the promising perspectives of the field [Be05][KO04]. However, it is still on a very early research stage and may be considered as 'novel'.

Collection of eye movement samples is a relatively challenging task, since it requires specialized devices (eye trackers) and carefully planned experiments to ensure the correct recording of the signals. Until recently, due to lack of publicly available datasets, every research team that developed a new scheme for the extraction of biometric features from the eye movements needed to conduct its own experiment and construct its own dataset. In this way though, a comparison among different methods that analyze eye movements for biometric purposes was very difficult to occur. The First Eye Movement Verification and Identification Competition (EMVIC) organized in 2012 as an official BTAS conference competition [KKK12] was the first to establish a common environment for the comparison of different approaches for the identification of individuals on the basis of their eye movements. The organizers prepared four different datasets of eye movements collected with different stimuli patterns and different eye trackers. There were about 50 competitors with over 500 separate submissions. An oddity that arouse during the EMVIC was that the identification results were inconsistent for the different prepared datasets. For datasets A and B the identification rates were better than 90% whereas for datasets C and D the best results approximated 60%. A question arose as to the reasons of such differences. In [KKK12] authors suggested several possible reasons. The first one was a number of recordings per person, because there were only 4 recordings per individual in datasets C and D and up to 100 recordings per individual in datasets A and B. The second reason was that data from both eyes were available for datasets A and B and data for one eye only were available for datasets C and D. The last two reasons were calibration and recording patterns impacts. It was pointed out that data in datasets A and B were not calibrated and most data were gathered using very close time proximity. Regarding the first reason, it is quite obvious that a bigger pool of samples per individual allows for a more complete evaluation of the similarities in the characteristics extracted from the experimental subjects. Detailed analysis regarding the second reason was given in [REF12][Ng12]. In our study we focus on the two latter reasons investigating how calibration, data quality and time proximity between samples may influence the final results.

3 Data preparation

The main objective during our experiments was to examine the influence of two different aspects on the final classification results: a) preprocessing of raw eye movements signals in order to improve samples quality, and b) temporal proximity of the recorded samples during the experiments. For this reason we have constructed several different subsets from a dataset of eye movements and used them with different algorithms in order to perform biometrical identification.

3.1 Dataset

The dataset on which we conducted our research was the publicly available dataset B from the EMVIC 2012 originally published in [BO05]. It consisted of 4168 samples taken from 75 subjects. The data were collected within a period of 9 consecutive months. The reason that led on the adoption of dataset B was the relatively large amount of data it provided. This made possible the extraction of different subsets of samples from the whole dataset and the application of different methodologies on these subsets in order to evaluate the impact of the inspected parameters on the identification performances.

3.2 Preprocessing of raw eye movements signals

During the recording phase of dataset B no calibration procedure occurred, so we decided to apply a post-calibration scheme and in the sequence a cleaning algorithm on the samples and inspect the impact of these data processing procedures on the final performance. Three groups of samples were created: *raw samples dataset*, which consisted of the samples in the exact form that they were recorded, i.e. without any calibration or cleaning. An analysis of the samples revealed that the quality of the data was relatively low and the amplitudes of signals may differ significantly among samples. Due to the nature of the stimulus it was easy to identify required fixation locations (RFLs) in the samples in every moment of registration [HH02]. This allowed us to perform a post-calibration procedure for the recorded raw signals. The post-calibration of a sample was implemented using the algorithm available in [Ka04]. The samples calibrated using the algorithm described above formed a *calibrated samples dataset*. Although for many samples the calibration was successful, there were cases where the resulting signal after the calibration was very noisy. To remove such bad conditioned samples a rejection threshold was added to the calibration algorithm. Signal levels calculated for different RFLs were compared and if the difference for two levels was lower than the rejection threshold, the sample was removed. The samples that passed that rejection test were used for the formation of the third group of samples, the *cleaned samples dataset*.

3.3 Temporal proximity of the recorded samples

The second parameter that was investigated regards the degree to which the time proximity among the recorded samples may influence the identification rates. Four different groups of datasets were constructed, each one corresponding to a different time interval between sample recordings. In order to perform a comprehensive evaluation procedure every dataset was created containing the same number of samples and the same distribution of subject identifiers, according to algorithm described in [Ka13].

The datasets denominated as ‘no interval’ consisted of samples with no minimal time separation condition. Since the samples in dataset B were mostly taken sequentially (from 1 to 15 samples per experiment) these datasets consisted mainly of samples taken at less than one minute’s intervals. Datasets denominated as ‘10 min’ were constructed with a minimal interval of 10 minutes between two samples taken from the same person. Similarly, for ‘1 h’ datasets the minimal interval between two samples was one hour and for ‘1 d’ datasets it was one day. A total of twelve separate datasets were created (four datasets for each of three groups of samples). Every dataset consisted of 275 samples taken from 38 individuals (38 classes), offering a sufficient amount of data for the evaluation of the inspected parameters significance on the identification performances of the employed classification methods.

4 Results

The datasets that were created with the preparation process described in Section 3 were employed in an ensemble of classification experiments. In our research we have used four different methods in order to classify the collected data and perform biometric identification via the eye movements. The first two methods are a spectral processing scheme (MEL) introduced in [Ng12] and a graph based approach (GRAPH) [REF12]. Both methodologies were developed specifically for the extraction of biometric traits from eye movement data and were successfully used during the EMVIC 2012. On the other hand we employed two universal methods, the J48 (Java implementation of C4.5) and the Random Forest (RF) algorithms, that are generally known for their efficiency in classification tasks. In this section the identification results are presented and conclusions are drawn with regard to the influence of the investigated parameters on the biometric performance of the benchmark methods.

4.1 Influence of data quality

The first factor that will be analyzed concerns the influence of the calibration and the cleaning procedure on the identification performances. The averaged results for the different methods that were used may be observed in Figure 1. For the three processing scenarios (raw, calibrated and cleaned) the rank-1 accuracy - i.e. number of correctly identified samples to the overall number of samples - is demonstrated for every method.

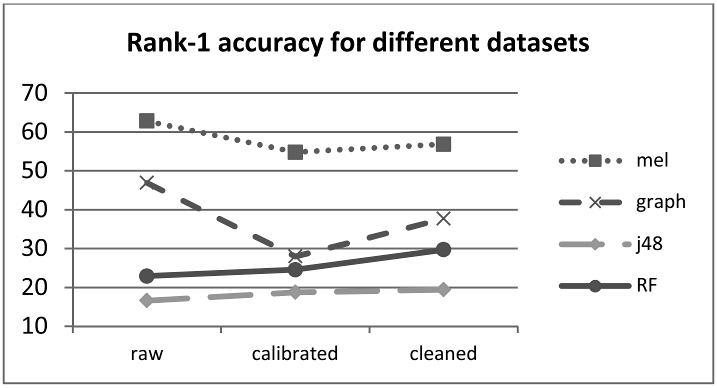

Figure 1: Average rank-1 identification accuracy for every method for the three types of samples processing (raw, calibrated and cleaned).

A close inspection of the identification rates shows that the *calibration* procedure significantly worsens results for the GRAPH method. The results are also worse (but not significantly) for MEL method, which may be surprising since this method works in the frequency domain. *Calibration* does not affect classic methods J48 and RF. In fact, in most cases results for calibrated data are even better although the difference is not significant. A possible explanation of the observed results may be the following: Samples in *raw* format have different amplitudes. Those amplitudes may somehow enclose individually distinctive information. For instance it may be connected with the shape of the face or the way a person wears eye tracking equipment.

Regarding now the data cleaning procedure, we can see that it has no significant effect on MEL and J48 methods. It improves results for RF but the improvement is not significant (p>0.05). Yet, it significantly improves results for GRAPH method. This leads to the conclusion that the framework in which the data are processed by the latter method makes it more sensitive to low quality data.

4.2 Influence of time intervals

The other analyzed parameter during our experiments regarded the influence of time intervals between samples recording on the identification performance. In Table 1 we can observe the effect of time-interval between the eye movement recordings on the performance of each method, for the raw samples (i.e. without calibration and cleaning).

Table 1. Results for raw datasets

interval	MEL	GRAPH	J48	RF
No	86.2	76.4	26.5	37.8
10 min	66.9	41.2	17.1	24.4
1 h	49.8	37.1	12.4	16.7
1 d	48.4	33.0	10.5	12.7

As it can be observed, time interval between samples has a significant impact on the final results. For all methods, a longer interval gradually worsens the identification rates. The difference is the most significant during the transition from no-interval to 10 minutes interval. The results for datasets from the second group (calibrated samples) are presented in Table 2. A similar behavior between time-interval and accuracy exists in this case too.

Table 2. Results for calibrated datasets

interval	MEL	GRAPH	J48	RF
No	73.6	46.3	26.7	37.7
10 min	60.1	28.0	20.5	26.7
1 h	45.1	20.6	14.7	17.6
1 d	40.3	17.0	13.2	16.1

Samples in the third group were calibrated and cleaned to remove outliers - i.e. samples with low quality. The results for datasets from the third group are presented in Table 3. We may observe the same negative correlation between time-interval and accuracy.

Table 3. Results for calibrated and cleaned datasets

interval	MEL	GRAPH	J48	RF
No	74.6	53.8	24.6	41.0
10 min	62.9	39.6	19.1	30.1
1 h	49.2	31.4	16.8	27.0
1 d	40.6	26.0	17.2	20.7

The results presented above lead to the conclusion that the time recording pattern (that is time between subsequent samples) seems to have a great impact on the identification results. The impact is more significant for short intervals (i.e. between no interval and 10 minutes interval) and is not so important (yet visible) for longer intervals (i.e. between 1 hour and 1 day). Consequently, the time interval during the recording of the samples is a factor that should be always taken into consideration during the inspection of the identification rates of any behavioral biometric identification method, in order to have a more complete view regarding the operational framework of the method. It should also be always taken into account that several factors as the participants' attitude, mood or physical condition may have an influence on the measurements.

5 Conclusions

In our experiments we have tested two very important parameters that affect the classification accuracy in behavioral biometrics methods, the quality of the samples and the time interval between consequent enrollments for every subject. Regarding the impact of data quality, it was observed that when the samples are employed in raw format - without any calibration and cleaning - it is probable that some biases may arise which can influence the results. Such phenomena may artificially change the resulting rates, so they should be considered during the setting of an experiment, or during the samples processing procedure. Another important factor that should always be taken into consideration is the time interval between the samples recordings. Our recommendation is that data taken in short term series may be used for classification experiments only when samples from the same series are not mixed in both training and testing sets. Unlike in most physiological biometrics, behavioral biometrics experiments are not mutually independent and we observed that the dependency between samples is inversely proportional to the time interval between the samples. In our opinion this dependency deserves more investigation in future.

Finally, under the light of the presented experimental findings we offer a list of suggestions regarding possible meta-information that may be considered during reporting research experiments results that concern behavioral biometrics: General information about the data: (1) number of samples, (2) number of subjects and additional information: (3) distribution of number of samples per subject, (4) minimal time between subsequent samples, (5) minimal time between subsequent samples of one subject. If the dataset is designed to publication, it is preferred to publish it without any preprocessing. If only classification results are published, it is important to add information about: (6) algorithm used for data calibration (if applied), (7) algorithm used to remove samples with low quality.

Our research shows that all parameters mentioned above may have a strong impact on classification results and therefore should not be omitted. As all experiments during the research were performed only for one modality - eye movements – we believe that our findings could be generalized to all behavioral methods. Naturally, the latter statement deserves further study.

Acknowledgements. Authors would like to thank Nguyen Viet Cuong for his contribution in performing Mel-Frequency classification and Oleg Komogortsev for his initial inspiration.

References

[Be05] Bednarik, R.; Kinnunen, T.; Mihaila, A.; Fränti, P.: Eye-movements as a biometric, In 14 Scandinavian Conference on Image Analysis, Lecture Notes in Computer Science, Springer-Verlag, vol. 3540, pp. 780-789, 2005.

[BO05] Brzeski, R.; Ober, J.: The biometrical system of the authentication realized on the ground of the movement of the eye, Techniki Komputerowe, Biuletyn Informacyjny IMM, 2005.

[Ca11] Cappelli, R.; Ferrara, M.; Maltoni, D.; Turroni, F.: Fingerprint Verification Competition at IJCB2011 Proceedings of International Joint Conference of Biometrics, 2011.

[De95] Deane, F.; Henderson, R.; Mahar, D.; Saliba, A.:. Theoretical examination of the effects of anxiety and electronic performance monitoring on behavioural biometric security systems, Interacting with Computers vol 7 no 4 (199.5) 395411, 1995

[Ga05] Gamassi, M.; Lazzaroni, M.; Misino, M.; Piuri, V.; Sana, D.; Scotti, F.: Quality assessment of biometric systems: a comprehensive perspective based on accuracy and performance measurement, IEEE Transactions on Instrumentation and Measurement, 2005.

[GT07] Grother, P.; Tabassi, E.: Performance of Biometric Quality Measures, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007.

[HH02] Hornof, A. J.; Halverson T.: Cleaning up systematic error in eye tracking data by using required fixation locations. In Behavior Research Methods, Instruments, and Computers, 34. 2002.

[HK06] Hicklin, A.; Khanna, R.: The role of data quality in biometric systems. Technical Report, Mitretek Systems, 2006.

[Ja00] Jain, A.; Hong, L. et al.: Biometric identification. Communications ACM 43(2): 90-98, 2000.

[Ka04] Kasprowski, P.: Human identification using eye movements. Doctoral thesis. Silesian Unversity of Technology, Poland, 2004.

[Ka13] Kasprowski, P.: The Impact of Temporal Proximity between Samples on Eye Movement Biometric Identification. Computer Information Systems and Industrial Management. LNCS 8104, Springer-Verlag, 2013.

[KKK12] Kasprowski, P.; Komogortsev, O. V.; Karpov, A.: First Eye Movement Verification and Identification Competition at BTAS 2012. IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS), 2012.

[KO04] Kasprowski, P.; Ober. J.: Eye Movement in Biometrics, Proceedings of Biometric Authentication Workshop, European Conference on Computer Vision in Prague 2004, LNCS 3087, Springer-Verlag., 2004.

[MW02] Mansfield, A. J.; Wayman J. L.: Best practices in testing and reporting performance of biometric devices, Teddington, Middlesex, UK: Centre for Mathematics and Scientific Computing, National Physical Laboratory, 2002.

[Ng12] Nguyen, V. C.; Vu, D.; Lam S.; Tung, H.: Mel-frequency Cepstral Coefficients for Eye Movement Identification. IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2012).

[Or03] Ortega-Garcia, J.; Fierrez-Aguilar, J.; Simon, D. at al.: MCYT baseline corpus: a bimodal biometric database. IEE Proc.-Vis. Image Signal Process., Vol. 150, No. 6, December 2003.

[REF12] Rigas, I.; Economou, G.; Fotopoulos, S.: Human eye movements as a trait for biometrical identification. IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 217-222, 2012.

[Wa02] Wayman, J. L.: Technical testing and evaluation of biometric identification devices, Biometrics, 2002; 345-368.