The Impact of Temporal Proximity between Samples on Eye Movement Biometric Identification

Paweł Kasprowski

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

Abstract. Eye movements identification is an interesting alternative to other biometric identification methods. It compiles both physiological and behavioral aspects and therefore it is difficult to forge. However, the main obstacle to popularize this methodology is lack of general recommendations considering eye movement biometrics experiments. Another problem is lack of commonly available databases of eye movements. Different authors present their methodologies using their own datasets of samples recorded with different devices and scenarios. It excludes possibility to compare different approaches. It is obvious that the way the samples were recorded influences the overall results. This work tries to investigate how one of the elements – temporal proximity between subsequent measurements – influences the identification results. A dataset of 2556 eye movement recordings collected for over 5 months was used as the basis of analyses. The main purpose of the paper is to identify the impact of sampling and classification scenarios on the overall identification results and to recommend scenarios for creation of future datasets.

Keywords: eye movement biometrics, behavioral biometrics, classification.

1 Introduction

The main problem of visual perception is that eyes register scene with uneven acuity. Only the part of the scene that falls on the fovea – region in the middle of the retina – is seen with correct sharpness. All other regions of retina are able to register only contours and fast movements. Therefore, eye movements are very important for correct recognition of objects in visual field. That is why eye movements are one of the fastest and the most accurate movements of a human being [8].

Eye movements may be divided into voluntary and involuntary. Voluntary eye movements are the effect of our will – we want to look at something. Involuntary eye movement is reflex action, automatic response to some stimulus, for instance sudden movement near the edge of vision. Both movements have physiological aspects but also depend on our previous knowledge or experience – having also important behavioral elements. That is why human identification using eye movements should be classified as behavioral biometrics [17][19].

K. Saeed et al. (Eds.): CISIM 2013, LNCS 8104, pp. 77–87, 2013.

1.1 Human Identification Using Eye Movements

Eye movements are yet another possibility to perform human identification. The idea that eye movements may be used for human identification is about 10 years old [16][20]. There have been several publications showing that the method is promising, however it is still on the very early research stage because collecting eye movements' data is difficult and eye movement capturing devices (eye trackers) are still relatively expensive. That is why in most cases the datasets collected by researchers are not publicly available.

Most of the eye movements recording experiments have used a 'jumping point' pattern originally introduced in [16]. In such kind of experiment the stimulus is forcing eye movements - the examined individuals should follow the point on the screen with their eyes. Such a recording is easy to analyze, because it requires that fixations and saccades happen in specific moments. However, there are several interesting experiments with different scenarios, including faces observation [25] or text reading [11]. There is also an attempt to perform identification without any information about a stimulus [18].

The problem common for all biometric methods using behavioral traits is so called *learning effect* [13]. When using the same stimulus for several times the person familiarizes with it and eye movements tend to become automatic. It is for instance clearly visible for texts – eye movements of a person reading the text that she already knows are very different from eye movements of the person reading a text new for her [24]. Such kind of reading is therefore often called *skimming*.

The learning effect is especially visible for very short intervals. A person that completes the same task for several times in very short period tends to "learn" the task and the movements (eye movements in our example) are becoming similar to each other. The effect of similarity between subsequent experiments is stronger for shorter periods and becomes invisible for very long periods (because human body "forgets" the task).

In our paper we tried to investigate how the interval between subsequent experiments performed by the same person influences the identification rates. To our best knowledge it is the first paper that analyses that aspect of eye movements' biometrics, however the problem has been already introduced in [15].

1.2 Eye Movement Verification and Identification Competition

There are several methods to analyze eye movements, but until quite recently it has been difficult to compare them due to lack of publicly available datasets (like for fingerprints [5] or faces [22]).

The First Eye Movement Verification and Identification Competition (EMVIC) organized in 2012 as an official BTAS conference competition was the first opportunity to compare different approaches [15]. The aim of the competition was to correctly identify individuals on the basis of their eye movements. The organizers prepared four different datasets of eye movements collected with different stimuli and different eye trackers (denominated A, B, C and D).

All datasets were divided into two parts:

- Training set, containing labeled samples;
- Testing set, containing samples with hidden labels.

The aim of the competitors was to build their classification models using labeled samples and then try to use those models to classify unlabeled samples from the testing set. There were about 50 competitors with over 500 separate submissions.

The main problem of EMVIC was that the results were inconsistent for different datasets. For datasets A and B the identification accuracies were better than 90% and for datasets C and D the best accuracies approximated 60%. The question arose as to the reasons of such differences.

One of the obvious reasons was higher number of samples per person for datasets A and B. Another reason could be binocular data in A and B (which was already studied in [26] and [21]). In [15] authors suggested also that low quality of data in A and B could be the reason of better performance. In this paper we investigate another of the possible reasons: impact of proximity between consecutive samples.

As eye movements measurement has very important behavioral aspect, we may expect that measurements taken in short periods of time may share some common information that is unwanted for identification purposes. For instance it has been proven that the person's attitude may highly influence eye movements. If a person is angry or amused, eye movement patterns are different than for the same person in neutral state [7]. Similarly, it is possible to find out the level of tiredness by examining eyes reaction to salience regions of images [27][12][28].

Other obvious time-dependent factors which are not directly connected with human properties but may influence the measurement are e. g. lightning conditions or existence of devices which may interfere with eye tracker like cellular phones, computers etc.

1.3 Dataset

To check the impact of short term learning effect we decided to perform experiments that would check if temporal differences in the dataset might change the overall accuracy results. We used a dataset of 2556 eye samples collected for over 5 months. There were 61 different subjects under the test with uneven distribution of number of samples (from 4 to 129 samples per subject). The dataset was originally a part of dataset B from EMVIC. Samples were taken with 250Hz frequency using Ober2 eye tracker. It was a jumping point on 2x2 matrix used as stimulus. One experiment lasted for 8128 ms and the stimulus was exactly the same for every experiment (Fig. 1). Every sample consisted of 2048 measurements of horizontal position of the left eye and 2048 measurements of vertical position of the left eye giving 4096 attributes. Additionally every sample had two properties: timestamp of measurement in seconds (t) and subject id (id).

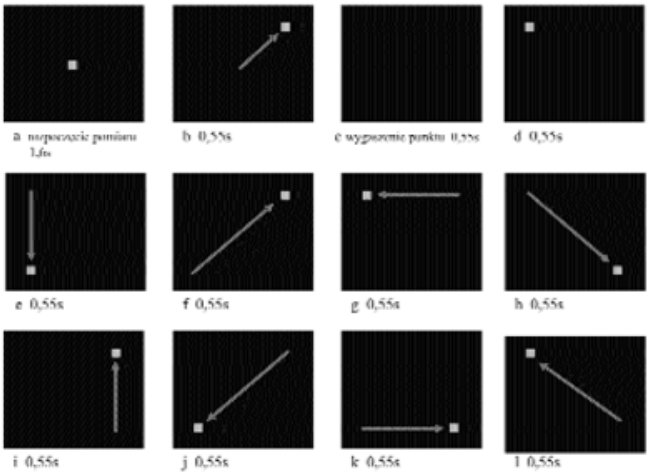

Fig. 1. Graphical representation of dataset [15]

2 Calculation of Time Interval Influence

To check the theory that time interval between samples influences the accuracy results several datasets were created with the same number of samples and the same distribution of samples per subjects. The only element that differed datasets was the minimal time interval between samples belonging to the same person.

2.1 Data Preparation

We started with creating a dataset for which a minimal time interval (MTI) between subsequent measurements of the same person was one week (604,800 seconds). We used original samples from the full dataset (D) and created a new dataset (W) by taking only samples fulfilling the interval condition. Algorithm filtering dataset D is shown below, where *t(s)* is the timestamp of measurement for sample *s*.

```
sort all samples in D by timestamp
foreach(id: ids from D)
  t_last = 0
  foreach(s: samples with given id from D)
    If t(s)-t_last > MTI
      add sample s to dataset W
      t_last = t(s)
```

The new dataset W consisted of 222 samples belonging to 37 subjects. The average number of samples per person was 6 with minimal value equal to 4 and maximal value equal to 11.

The next step was creation of other datasets. The property that differed datasets was a minimal time interval (MTI) between subsequent measurements of the same person. Every dataset consisted of the same number of samples for each person and was created using a subset of original samples from dataset D. The algorithm to build datasets was almost identical to the presented above. An additional parameter was the way the samples were initially sorted. There were two possibilities: start from the beginning of the dataset (with samples recorded at the beginning of the experiment) or start from the end of the dataset (with samples recorded as the last samples during the whole experiment). Because we wanted to see if it influences results, we decided to use both methods and store samples in datasets denominated as F for the oldest samples and R for the newest.

The above procedure created two datasets ('F' and 'R') for every interval. There were seven different values of MTI used as presented in Table 1. The columns 'real minimal/maximal intervals' show actual intervals for subsequent trials found in datasets.

Table 1. Minimal time intervals used for experiments

index	MTI	real minimal interval	real maximal interval
0	0 (no minimal interval)	11s	13d, 21h, 31min, 26s
1	1 minute	61s	46d, 3min, 5s
2	10 minutes	13min, 25s	46d, 3min, 5s
3	1 hour	1h, 1s	57d, 19h, 44min, 45s
4	6 hours	6h, 10min, 11s	66d, 23h, 13min, 9s
5	1 day	24h, 33min, 42 s	66d, 23h, 13min, 9s
6	1 week	7d, 6s	66d, 23h, 13min, 9s

The main idea of the paper was that identification results were dependent on minimal time interval between samples. It was assumed that samples taken in shorter intervals have some additional (usually unwanted) time related information that could improve classification results. The datasets were examined using four different classic classification algorithms, namely:

- J48 (Java version of C45 algorithm [23]),
- Random Forest [3],
- Naïve Bayes
- SVM (using Sequential Minimal Optimization algorithm) [29].

Every dataset was validated using standard 10-fold cross-validation method. The result for every dataset-algorithm pair was then stored as accuracy value. Accuracy is the number of correctly classified samples to the overall number of samples. Because it was 37 different classes, probability of random guess was less than 3% and therefore accuracy seemed to be a good and sufficient measure.

It is very important to emphasize that all algorithms were used with standard parameters [10] without any optimizations towards results improvements. As the main purpose of the paper was to compare different datasets in the same environment (and not to obtain the best result), additional parameters tuning could introduce some biases. Nevertheless, the results are quite good as for identification (one-to-many) task.

2.2 Results

The results of the experiment are presented in Table 2.

Table 2. Accuracy of each classification method for every dataset

dataset	nb	j48	rf	smo
0F	22,97	20,27	28,38	52,25
0R	28,37	21,17	37,39	54,04
1F	18,46	13,51	18,02	31,53
1R	26,12	17,56	24,32	44,59
2F	17,11	11,26	23,87	33,78
2R	25,67	20,72	22,97	48,65
3F	17,56	12,16	22,97	34,23
3R	22,97	22,97	21,17	45,95
4F	18,46	13,96	18,47	32,43
4R	16,66	12,16	18,47	33,33
5F	18,01	13,51	20,27	31,08
5R	18,01	16,21	14,41	36,94
6F	12,16	12,61	15,31	29,73
6R	12,16	13,06	16,67	29,73

Datasets with F suffix were created by taking samples starting from the earliest while datasets with R suffix were created by taking samples from the last experiments (as it was described in the previous section). Every dataset had the same number of samples (222) and the same distribution of samples among 37 different subjects.

What can be seen clearly from the results is strong negative correlation between the interval and accuracy (-0.61). Because so called “forgetting curve” [9] is considered to be non-linear we also calculated correlation of accuracy with logarithm of the interval and obtained the correlation equal to -0.94. Fig 2 shows logarithmic regression of the average results with coefficient of determination (R^2) equal to 0.8637. For linear regression coefficient of determination was 0.8088 and for exponential regression was 0.8433 so logarithmic trend line was chosen as the best fitting option.

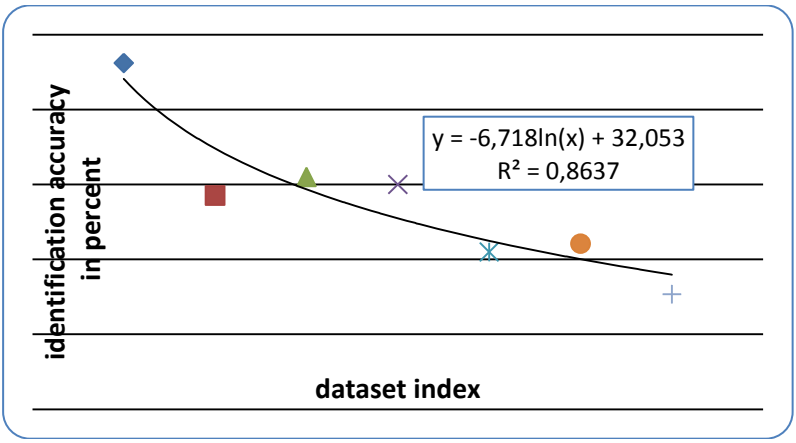

Fig. 2. Averaged accuracy results with logarithmic trend line

The accuracies of datasets classification were averaged for all datasets with the same interval and all classification methods. When calculating mutual significance of differences between these results it occurred that the only significant difference may be found between 0 and 1 minute interval (p=0.02). It means that memory effect is clearly visible only for very short intervals.

Another comparison was performed between samples of type 'F' (i.e. first samples of the specific person) and samples of type 'R' (i.e. last samples of the person). The hypothesis was that samples taken later - when a person is already familiarized with stimulus - will be more stable and therefore easier to classify. Indeed, average accuracy for 'R' datasets was better for every classification method. However, the differences were not significant, with the highest significance for SVM method (p=0.068).

To see if the results are stable for different signal conversions we repeated the same classification experiments on datasets converted using different algorithms previously used in eye movement biometric identification [1][4][11][14][15][16][18][21][25][20][13]. The results were similar, always showing negative correlation, however for some conversions the correlation was not strong.

Table 3. Correlation of classification accuracy and minimal time interval between samples for different signal conversions (*time* means correlation to time in seconds, *log* is correlation to logarithm of time in seconds)

applied conversion	time	log(time)
fourier spectrum	-0.42	-0.91
cepstrum	-0.7	-0.81
first derivate (velocity)	-0.22	-0.74
second derivate (acceleration)	-0.48	-0.92
direction (in radians)	-0.8	-0.84
wavelet transform (DWT)	-0.46	-0.86
high pass filter	-0.56	-0.93
low pass filter	-0.63	-0.94

3 Sessions Analyzes

The experiment presented in the previous section showed clearly that samples taken in short intervals should not be mixed in training and testing sets for classification.

On the other hand, closer look into the dataset that was used for experiments revealed that in most cases samples of the same person were recorded in series. It is natural and it is probably a common strategy for every biometric experiment, because with this scenario only one equipment setup is required to obtain several samples. However, as it was proven in the previous section, samples taken in the same session are not independent. Therefore, we decided to divide the original dataset into sessions and check how it influences classification results.

The *session* was defined as a set of samples taken from the same person with minimal interval between two samples less than 10 minutes. Preprocessing algorithm found 685 sessions in the dataset. The number of sessions per subject differed from 1 to 26 sessions with average number of sessions equal to 11. Every session consisted of one to eight samples (see Fig. 3).

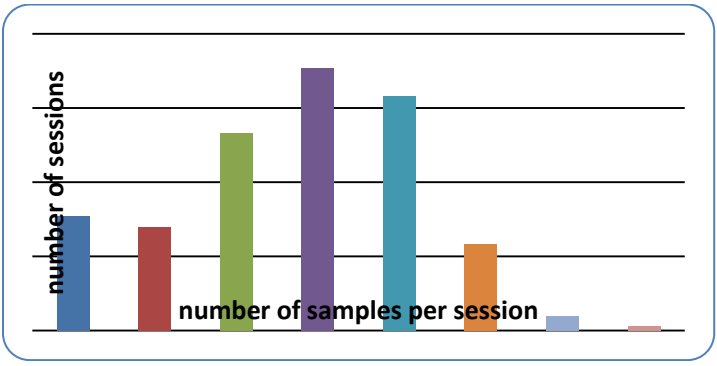

Fig. 3. Histogram of number of samples per session

To check identification results using 10-fold cross-validation and to obtain reliable results it was necessary to remove samples of subjects for which number of sessions was too low. Therefore samples of all subjects with less than 10 sessions were removed from dataset. The reduced dataset consisted of 2195 samples from 29 subjects divided into 567 sessions.

Because samples from the same session were considered to be dependent there were three classification experiments proposed:

- Experiment using only first samples from each session (referred as *first sample* in Table 4).
- Experiment using only last samples from each session (*last sample*).
- Experiment using all samples from dataset but with folding algorithm that doesn't divide samples from the same session to different folds (*all samples*).

Contrary to cross validations used in experiment described in Section 2, where folds were created by stratifying basing on number of samples per subject, this time stratification was done basing on sessions. It means that all samples from the same session had to be in the same fold.

Table 4. Accuracies of classification for three datasets

method	first sample	last sample	all samples
J48	23,27	21,02	29,33
NB	28,39	25,80	27,94
RF	34,41	29,52	40,58
SMO	50,09	52,12	64,83
average	**34,04**	**32,12**	**40,67**

Table 4 shows that there are no significant differences between datasets build from first and last samples from the session. However, when all samples from the session were taken, it significantly improved results. It must be remembered that the latter classification used much more samples both for training and testing (2195 versus 567). It shows that collecting samples in series is not generally a bad idea, but care must be taken how samples from the same session are used.

4 Conclusions

The results of analyzes presented in the paper clearly show that the data collecting scenario may significantly influence the overall results and classification possibilities for a dataset. Especially time related factors were carefully studied and impact of so called memory effect was analyzed.

All calculations used only eye movements datasets but it may be assumed that the conclusions could be extended to other behavioral biometric experiments.

Basing on our findings we advise that every behavioral biometric sample should be stored together with information about the exact timestamp when it was collected.

It is possible to collect more than one sample during one session with subject but these samples should never be mixed in training and testing set when evaluating performance. Additionally, we showed that using all samples collected during one session in the training set improves the overall performance of the system. Even if samples from the same session were considered dependent, multiplying the number of samples would give effect similar to bootstrap samples used in bagging algorithm [2].

References

1. Bednarik, R., Kinnunen, T., Mihaila, A., Fränti, P.: Eye-movements as a biometric. In: Kalviainen, H., Parkkinen, J., Kaarna, A. (eds.) SCIA 2005. LNCS, vol. 3540, pp. 780–789. Springer, Heidelberg (2005)

2. Breiman, L.: Bagging predictors. Machine Learning 24(2) (1996)
3. Breiman, L.: Random Forests. Machine Learning 45(1), 5–32 (2001)
4. 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)
5. Cappelli, R., Ferrara, M., Maltoni, D., Turroni, F.: Fingerprint Verification Competition at IJCB 2011 Proceedings of International Joint Conference of Biometrics (2011)
6. Chen, Y., Dass, S.C., Jain, A.K.: Localized iris image quality using 2-d wavelets. Advances in Biometrics (2005)
7. 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 7(4), 395–411 (1995)
8. Duchowski, A.: Eye Tracking Methodology. Theory and Practice. Springer, London (2003)
9. Ebbinghaus, H.: Memory: A contribution to experimental psychology. Teachers college, Columbia university (1913)
10. Hall, M., et al.: The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter 11(1), 10–18 (2009)
11. Holland, C., Komogortsev, O.V.: Biometric Identification via Eye Movement Scanpaths in Reading. In: IEEE International Joint Conference on Biometrics, IJCB (2011)
12. Ji, Q., Zhu, Z., Lan, P.: Real-time nonintrusive monitoring and prediction of driver fatigue. IEEE Transactions on Vehicular Technology 53(4), 1052–1068 (2004)
13. Kasprowski, P., Ober, J.: Enhancing eye movement based biometric identification method by using voting classifiers. In: SPIE Defence & Security Symposium, SPIE Proceedings, Orlando, Florida (2005)
14. Kasprowski, P.: Human identification using eye movements. Doctoral thesis. Silesian University of Technology, Poland (2004)
15. 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)
16. Kasprowski, P., Ober, J.: Eye Movements in Biometrics. In: Maltoni, D., Jain, A.K. (eds.) BioAW 2004. LNCS, vol. 3087, pp. 248–258. Springer, Heidelberg (2004)
17. Keesey, U.T.: Effects of involuntary eye movements on visual acuity. JOSA (1960)
18. Kinnunen, T., Sedlak, F., Bednarik, R.: Towards task-independent person authentication using eye movement signals. In: Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications. ACM, New York (2010)
19. Leigh, R.J., Zee, D.S.: The Neurology of Eye Movements. Oxford University Press (2006)
20. Maeder, A.J., Clinton, B.F.: A visual attention approach to personal identification (2003)
21. Nguyen, V.C., Vu, D., Lam, S., Tung, H.: Mel-frequency Cepstral Coefficients for Eye Movement Identification. In: IEEE International Conference on Tools with Artificial Intelligence, ICTAI (2012)
22. Ortega-Garcia, J., Fierrez-Aguilar, J., Simon, D., et al.: MCYT baseline corpus: a bimodal biometric database. IEE Proc.-Vis. Image Signal Process. 150(6) (2003)
23. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers (1993)
24. Rayner, K.: Eye movements in reading and information processing: 20 years of research. Psychological Bulletin 124(3), 372 (1998)
25. Rigas, I., Economou, G., Fotopoulos, S.: Biometric identification based on the eye movements and graph matching techniques. Pattern Recognition Letters 33(6), 786–792 (2012) ISSN 0167-8655

26. Rigas, I., Economou, G., Fotopoulos, S.: Human eye movements as a trait for biometrical identification. In: IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 217–222 (2012)
27. Schleicher, R., et al.: Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired? Ergonomics 51(7), 982–1010 (2008)
28. Sodhi, M., Reimer, B., Llamazares, I.: Glance analysis of driver eye movements to evaluate distraction. Behavior Research Methods, Instruments, & Computers 34(4), 529–538 (2002)
29. Vapnik, V.N.: Statistical learning theory. Wiley Interscience (1998)