Abstract—This paper presents the results of the first eye movement verification and identification competition. The work provides background, discusses previous research, and describes the datasets and methods used in the competition. The results highlight the importance of very careful eye positional data capture to ensure meaningfulness of identification outcomes. The discussion about the metrics and scores that can assist in evaluation of the captured data quality is provided. Best identification results varied in the range from 58.6% to 97.7% depending on the dataset and methods employed for the identification. Additionally, this work discusses possible future directions of research in the eye movement-based biometrics domain.

Index Terms—eye movements, biometrics, competition

I. INTRODUCTION

There are several ways to use eye related information for biometric purposes, e.g., iris [4][26], face recognition [41], retina [29], periocular information [39]. One of the additional biometric modalities related to the eye is biometrics based on eye movements. This biometric modality was suggested approximately 10 years ago [11][38], however, relatively few publications were written on this topic so far. To facilitate research in this area we have decided to organize an eye movement biometric competition.

The competition provided a common ground in a form of several datasets to benchmark the eye movement biometric methods derived by the participants. Our subsequent work with the results and the datasets allowed providing recommendations related to the eye movement data collection, measuring eye movement quality, and deciding when to record samples from the subjects to ensure meaningfulness of current and future benchmarking results.

This paper is organized as follows: Section II provides information related to the eye movements and their origin, also the information about eye movement data recording and assessing its quality; Section III briefly describes previous work related to the eye movement biometrics; Section IV provides the details of the datasets employed in the competition; Section V describes the competition protocol; Section VI outlines the results; Section VII provides a discussion about results and related data quality; Section VIII follows with a conclusion and future work.

II. EYE MOVEMENTS

A. Overview

Human eyes differ substantially from a common digital camera [22]. One of the differences is a non-uniform picture quality across the visual field. Specifically, the fovea – the high visual acuity zone of human retina is just approximately two degrees of the visual angle [28]. The acuity of vision sharply drops outside of the fovea. The center of the fovea - foveola is the central part of the retina that provides highest visual acuity. Visual axis can be extended from the foveola to the object perceived by the eye, thus creating a point of regard also called a gaze point. Non-uniformity of human vision necessitates eye movements with a goal of capturing visual information from the surrounding world.

Among different eye movements exhibited by the Human Visual System (HVS) following two types are most relevant to this work: fixation – eye is stable toward the object of interest, saccade – rapid eye rotation between fixation points. Miniature eye movements such as tremor, drift, and micro saccades [42] are a part of a fixation and keep an eye in constant motion. Constant movement is necessary due to the motion sensing nature of the light perceiving cells, which require constant excitation to translate the light to the neuronal signal [22]. Artificial stabilization of the eye globe leads to the loss of vision. Saccades are the fastest movements in the human body with velocities reaching several hundred degrees per second [37]. The eye is blind during saccades.

Eye movements may be divided into voluntary and involuntary. Voluntary eye movements are result of our will, making it possible to control the focus of our attention. Involuntary eye movement is a reflexive action, automatic response to some stimulus, for instance sudden stimulus movement near the edge of vision.

Physiologically, the eye movements are made possible by a) oculomotor plant – eye globe, three pairs of extraocular muscles, and surrounding tissues [37], b) different brain areas and mechanisms are responsible for programming oculomotor plant [37]. Extensive description of the related anatomical structures and their functionality are beyond the scope of this
work.

B. Gaze Data Recording and Quality Assessment

Eye movements are recorded by a device called an eye tracker, which reports raw eye positional data at specified sampling frequency [22]. Main characteristics of the eye tracking equipment are: a) positional accuracy – the difference between the reported and the actual gaze point, b) precision – minimum amount of gaze shift detectable by the equipment, c) sampling frequency. Detailed survey of the different eye tracking approaches can be found here [23].

Currently, two main metrics are widely accepted as quality indicators of captured raw eye positional data: calibration error and data loss.

Calibration error is determined during a calibration procedure. Calibration procedure is very important in the eye tracking research. Its goal is to train eye tracking software to estimate eye gaze position for every eye image captured by the image sensor. This goal is achieved by a presentation of pre-set target points that are usually uniformly distributed on the visual screen and requesting a subject to look at these predefined gaze locations [22]. Subsequently, when actual recording takes place, the gaze locations that fall outside of the initial calibration points are interpolated by various algorithms [23]. Calibration error indicates the average positional difference between the coordinates of the pre-set calibration points and the coordinates of the estimated gaze locations for those points. Calibration error varies depending on a subject and experiment setup. Calibration error is expected to be close to equipment’s positional accuracy, however, quite often the calibration error can be several times larger than positional accuracy reported by an eye-tracking vendor. Please note that calibration error can be also termed as accuracy or positional accuracy in the eye tracking literature (e.g. [27][35]).

Data loss is the amount of gaze samples reported as invalid by an eye tracker. Data loss is usually caused by blinking, head movements, changes in stimulus or surrounding lighting, and squinting. In cases when gaze points fall outside of the recording boundaries (e.g., computer screen), which usually happens due to poor calibration, these gaze points are marked as invalid. Please note that not all eye trackers are capable of marking invalid gaze points. In such cases invalid gaze points should be found and marked by the experimenter to compute resulting data loss.

Usually smallest calibration error and data loss for the eye tracking systems are achieved when subject’s head is fixated by a chin rest at an optimal distance from the image sensor (usually 40-70cm.). However, modern advances in the table and head mounted eye tracking systems allow to collect acceptable data quality when the head is not fixated.

More detailed discussion about gaze data quality can be found here [27].

C. Classification of Captured Gaze Data Into Fixations and Saccades

Raw eye positional signal should be classified into fixations and saccades (and other eye movement types when stimulus properties are likely to invoke them) in cases when it is necessary to assess performance of the Human Visual System (HVS) or to employ eye movement characteristics for biometric purposes. Several algorithms exist for classification purposes [33]. Classification of the raw gaze data into fixation and saccades also allows assessing the meaningfulness of the captured data via behavior scores that are discussed next.

D. Behavior Scores

Behavior scores provide the capability to assess the performance of the HVS, which is represented by the results provided by the eye movement classification algorithms [35][33]. Behavior scores support the idea that the HVS performance of a normal person matches signal characteristics encoded in the stimulus. Cases where HVS performance drastically differs from the ideal response of a normal person might be indicative of pathologies (e.g., head trauma), poor quality of the recorded signal, or/and failed eye movement classification. Currently, behavior scores can be computed only for pulse-step or pulse-step-ramp stimulus (moving dot of light) with known characteristics [33][31].

Following behavior scores are employed for data quality control in this work: Fixation Quantitative Score (FQnS) – measures the amount of detected fixational behavior in response to a stimulus, Fixation Qualitative Score (FQIS) compares the spatial proximity of the classified fixation signal to the presented stimulus signal. The FQIS usually highly correlates with the calibration error. Saccade Quantitative Score (FQnS) – measures the amount of detected saccadic behavior in response to a stimulus. Detailed discussion on the behavior scores and their ideal values can be found here [33].

III. EYE MOVEMENT BIOMETRICS: PREVIOUS WORK

Related work in the eye movement biometrics field can be approximately divided into four general categories:

a) Use of the raw eye positional signal and its derivatives [11][1][12][15]. Standard techniques for feature extraction are used including first/second derivatives of the signal, Wavelet transform, Fourier transform, and other frequency related transformations of the signal. Frequently methods such as Principal Component Analysis are employed to reduce the number of features. Template matching is done using such algorithms as K Nearest Neighbors, Naive Bayes, C45 Decision Trees and Support Vector Machines.

b) Use of oculomotor plant characteristics (OPC) [17], where OPC are extracted with the help of a mathematical model of the eye and OPC templates are matched by statistical methods such as Hotelling’s T-square test;

c) Inference of brain control strategies, and mechanisms responsible for guidance of visual attention via analysis of complex eye movement patterns (CEM) [24][25]. In CEM approach features are represented by several individual and aggregated characteristics that are fundamental to HVS with template matching occurring via probabilistic and distance related approaches. It is possible to include in this category approach investigated by Rigas and colleagues [20], where Minimum Spanning Tree structures representing sequences of fixation/saccades and distances between these structures are
IV. DATASETS FOR THE COMPETITION

Two types of the datasets were employed for the competition: uncalibrated and calibrated. Uncalibrated datasets were collected with a goal of minimizing data capture time, i.e., without calibration procedure or equipment adjustments. Therefore, captured data quality could not be controlled. Calibrated datasets were captured with every precaution to obtain highest possible data quality, i.e., equipment was adjusted and re-adjusted when necessary to ensure minimum calibration error and data loss. Depending on a subject, necessary iterative equipment adjustments, subsequent calibration, and its verification could sometimes increase data capture times considerably. Table I provides technical details for each database. Next two sections provide additional details.

A. Uncalibrated Datasets

Dataset A. Step stimulus was presented in a form of a jumping dot interpolated on the 3x3 grid. The stimulus consisted of eleven dot position changes giving twelve consecutive dot positions. Subjects were given instructions to follow the dot. First dot appeared in the middle of the screen. After 1600 ms the dot in the middle disappeared and for 20 ms a screen was blank. Subsequently, a dot appeared in the upper right corner. The sequence continued until all locations of the 3x3 grid were visited. Dot movements on the grid were interspersed with dots presented at the central screen location. Figure 1 provides additional stimulus details. Figure 2 presents a sample from the recorded signal. Figure 3 presents a histogram of time intervals between the recordings.

Dataset B. Step stimulus was presented in a form of a jumping dot interpolated on the 2x2 grid. Dot’s position changed after every 550 ms. Additionally there was 550 ms "break" with no dot (black screen) in the middle of stimulation. Maximum of 10 recordings per day were conducted per subject. Figure 3 presents a histogram of time intervals between the recordings.

B. Calibrated Datasets

Dataset C. For each recording session the step stimulus was presented as a vertical jumping dot, consisting of a grey disc sized approximately 1 degree of visual angle with a small black point in the center. The dot performed 120 vertical jumps with amplitude of 20 degrees of the visual angle. At each spot the dot was displayed for 1s. The first two recordings for each subject were conducted during the same day with an interval of approximately 15 minutes; two more recordings were conducted one week later during a single day with an interval of approximately 15 minutes. Chin rest was employed to stabilize subjects’ heads during the recording. Figure 2 presents a sample from the recorded signal. Figure 3 presents a histogram of time intervals between the recordings.

Dataset D. For each recording session the dot appeared 100 times each time at random location with only requirement that at the end of the 100 appearances the spatial placement of the dots on the screen would be close to uniform. Other recording parameters are the same as for the Dataset C. Videos depicting stimulus for Datasets C and D can be found at www.emvic.org. Figure 3 presents a histogram of time intervals between the recordings. Datasets C and D are a part of a larger biometric database that can be downloaded here [32].
TABLE I. TECHNICAL DETAILS FOR EACH DATASET. BEHAVIOR SCORES FOR DATASETS A AND B WERE COMPUTED AFTER POST CALIBRATION [13]. PRIOR TO SCORE COMPUTATION SOME RECORDINGS IN DATASETS A AND B WITH VERY POOR DATA QUALITY WERE REMOVED. DATA LOSS FOR DATASETS A AND B WERE COMPUTED WHEN SIGNAL WAS DETECTED OUT OF SCREEN BOUNDARIES.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Dataset A</th>
<th>Dataset B</th>
<th>Dataset C</th>
<th>Dataset D</th>
</tr>
</thead>
<tbody>
<tr>
<td># of subjects</td>
<td>37</td>
<td>75</td>
<td>29</td>
<td>27</td>
</tr>
<tr>
<td>Total # of recordings</td>
<td>978</td>
<td>4168</td>
<td>116</td>
<td>108</td>
</tr>
<tr>
<td>Recordings per subject</td>
<td>4-158</td>
<td>5-172</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Recording duration</td>
<td>8.2s</td>
<td>8.2s</td>
<td>120s</td>
<td>100s</td>
</tr>
<tr>
<td>Maximum time between the first and the last recording</td>
<td>3 months</td>
<td>4 months</td>
<td>12 days</td>
<td>12 days</td>
</tr>
</tbody>
</table>

Experimental setup

Eye Tracker (ET) | Ober2 | Ober2 | Eye Link | Eye Link |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ET accuracy</td>
<td>N/A</td>
<td>N/A</td>
<td>0.25°-0.5°</td>
<td>0.25°-0.5°</td>
</tr>
<tr>
<td>ET precision</td>
<td>N/A</td>
<td>N/A</td>
<td>0.01°</td>
<td>0.01°</td>
</tr>
<tr>
<td>ET sampling rate</td>
<td>250Hz</td>
<td>250Hz</td>
<td>1000Hz</td>
<td>1000Hz</td>
</tr>
<tr>
<td>ET type</td>
<td>head mounted</td>
<td>head mounted</td>
<td>remote</td>
<td>remote</td>
</tr>
<tr>
<td>Chin rest</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Display size</td>
<td>17&quot;</td>
<td>17&quot;</td>
<td>30&quot;</td>
<td>30&quot;</td>
</tr>
<tr>
<td>Distance head to screen</td>
<td>700mm</td>
<td>700mm</td>
<td>685mm</td>
<td>685mm</td>
</tr>
</tbody>
</table>

Data Quality

| Average calibration error (SD) | N/A | N/A | 0.73° (0.39) | 0.75° (0.55) |
| Average Data Loss (SD)        | 9.06% (14.53) | 35.19% (25.45) | 2.88% (0.04) | 2.88% (0.04) |

Behavior scores

| Ideal_FQnS | 66% | 63% | 74% | 75% |
| FQnS (SD) | 19% (13.4) | 10% (9) | 59% (9.2) | 65% (8.2) |
| Ideal_FQIS | 0° | 0° | 0° | 0° |
| FQIS (SD) | 2.14° (0.55) | 2.11° (0.59) | 1.11° (0.42) | 1.38° (0.33) |
| Ideal_SQnS | 100% | 100% | 100% | 100% |
| SQnS (SD) | 116% (34) | 149% (50) | 108% (62) | 116% (31) |

C. Separation into Training and Testing Sets

Datasets A and B were separated into the training set that contained 65% of the recordings and the testing set that contained 35% of the recordings. The assignment of recordings to each set was made by random stratified sampling.

Datasets C and D were separated into the training set and testing sets via 50% / 50% split. The testing set contained the recordings that were performed during the first week of the experiments and thus each subject was represented by two recordings. The training set contained the recordings that were performed during the second week of experiments and contained remaining two recordings for each subject.

V. COMPetITION PROTOCOL

The aim of the competitors was to build their classification models using labeled recordings in the training sets and then try to use those models to classify unlabeled recordings in the testing sets.

In many similar competitions (e.g. Fingerprint Verification Competition [43]) competitors should send the organizers an application that is able to take given input data and produce output in the specified format. Usually, when ranking the performance of such applications both the execution time and the accuracy are taken into the consideration. However, we decided to simplify the submission process. Competitors had to send only a file in pre-defined format that contained the identity of each individual in the testing set. Submission format allowed for strong classification (only one identity for each recording) or weak (multiple identities each marked with a probability value).

The competition was broken into two parts: the main competition (www.emvic.org) that consisted of all four datasets and an additional competition on Kaggle web service (http://www.kaggle.com/c/emvic). The Kaggle's part was simplified and consisted of the dataset A only with a slightly different submission file format as required by the host.

The main competition required for every recording a list of probabilities that a given recording belongs to a specific subject (sid) in format sid:prob,sid:prob. The number of sid:prob pairs was not specified. It could be just one sid:1 for the strong classifier case or a list of all sids and probabilities in case of the weak classifier. We encouraged competitors to send sets of sid:prob pairs to enable the extraction of various performance parameters, however, very few participants sent this information.

The competition employed rank-1 identification accuracy benchmark marked as ACC1. It was computed as the number of records classified correctly to the whole number of records. Correct classification is when correct sid is marked by the highest probability.

For Kaggle competition the file format was slightly different, but contained the same information. Log loss metric was suggested by Kaggle project administration and was adopted as a performance benchmark.

Log loss is defined as:

\[
\logloss = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij} \log\left(y_{ij}^{^\wedge}\right)
\]

where N is the number of samples, M is the number of subjects, log is the natural logarithm, \(y_{ij}^{^\wedge}\) is the posterior probability that the jth subject generated the ith sample, and \(y_{ij}\) is the ground truth (\(y_{ij}=1\) means that the jth subject generated the ith sample, \(y_{ij}=0\) indicates otherwise).

Log loss metric guarantees that submissions in which correct sids are marked by the high probability (not necessary the highest one) get better score.

Kaggle competition participants were limited to two daily submissions. Table II indicates that substantial number of attempts (frequently more than 20) was required to obtain best results. To minimize over-fitting only 25% of the test dataset A (so called “public” part) was given for these trials.
Competitors sending their solutions were aware of their public score only. The final score was calculated using the whole testing dataset. Whole testing Dataset A frequently yielded worse log loss scores.

VI. COMPETITION RESULTS

There were overall 49 competitors in Kaggle hosted competition and 45 registered users in the main competition. There were 524 submissions sent in Kaggle competition and 106 in the main. Tables II and III present the summary of the results.

As a part of the competition participants filled out a survey where they had an opportunity to provide an account of methods and tools they employed to achieve their results.

Most of the participants (especially “kagglers”) treated the data in the individual recordings as a sequence of numbers and treated those sequences via general signal processing and data mining algorithms. Except raw eye positional signal only features related to the first and second derivate of the signal (i.e. velocity and acceleration) were considered. Extraction of fixations and saccades from raw eye positional signal and employment of any features related to those events was not reported. The only reported exception was signal parsing into segments via fixations.

Because the number of attributes extracted from the signal was substantial, most of the competitors used some techniques to reduce dimensionality. The most popular techniques involved dividing a recording into some subsets (typically 8 to 16) and summarizing these subsets by their characteristics, e.g., average velocity, average spatial location. Among others, methods such as Singular Value Decomposition or/and Principal Component Analyses led to a reasonable reduction of dimensionality. Among various techniques employed for the template matching SVM, Random Forest and k Nearest Neighbors methods were most popular and produced highest identification accuracy results.

<table>
<thead>
<tr>
<th>Team</th>
<th>Method</th>
<th>Entries</th>
<th>Public score</th>
<th>Final score</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRIG</td>
<td>kNN</td>
<td>28</td>
<td>0.23</td>
<td>0.23</td>
</tr>
<tr>
<td>Killian O.</td>
<td>Random Forest</td>
<td>18</td>
<td>0.12</td>
<td>0.37</td>
</tr>
<tr>
<td>Dorothy</td>
<td>Random Forest + LDA</td>
<td>24</td>
<td>0.18</td>
<td>0.48</td>
</tr>
<tr>
<td>zeon</td>
<td>n/a</td>
<td>24</td>
<td>0.33</td>
<td>0.52</td>
</tr>
<tr>
<td>GeLo</td>
<td>n/a</td>
<td>10</td>
<td>0.33</td>
<td>0.59</td>
</tr>
</tbody>
</table>

TABLE II. TOP FIVE RESULTS OF THE KAGGLE’S COMPETITION. EACH SCORE REPRESENTS LOG LOSS.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Methodology</th>
<th>ACC1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michal Hradiš, Brno University</td>
<td>2D histogram speed and direction, SVM</td>
<td>97.55%</td>
</tr>
<tr>
<td>Ioannis Rigas, University of Patras, Greece</td>
<td>Multivariate Wald-Wolfowitz test, kNN</td>
<td>96.63%</td>
</tr>
<tr>
<td>Nguyen Viet Cuong, National Univ of Singapore</td>
<td>Bayesian Network with Mel-frequency cepstral coefficients features</td>
<td>93.56%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dataset B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Michal Hradiš, Brno University</td>
<td>2D histogram speed and direction, SVM</td>
<td>95.11%</td>
</tr>
<tr>
<td>Ioannis Rigas, University of Patras, Greece</td>
<td>Multivariate Wald-Wolfowitz test, kNN</td>
<td>90.43%</td>
</tr>
<tr>
<td>Nguyen Viet Cuong, National Univ of Singapore</td>
<td>SVM with Mel-frequency cepstral coefficients features</td>
<td>90.43%</td>
</tr>
<tr>
<td>Dataset C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Michal Hradiš, Brno University</td>
<td>Nearest neighbor with X2 distance. 2D histograms on short windows - speed and direction.</td>
<td>58.62%</td>
</tr>
<tr>
<td>Nguyen Viet Cuong, National Univ of Singapore</td>
<td>SVM with Mel-frequency cepstral coefficients features + PCA</td>
<td>37.93%</td>
</tr>
<tr>
<td>Ioannis Rigas, University of Patras, Greece</td>
<td>Multivariate Wald-Wolfowitz test, kNN</td>
<td>25.86%</td>
</tr>
<tr>
<td>Dataset D</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Michal Hradiš, Brno University</td>
<td>Nearest neighbor with X2 distance. 2D histograms on short windows - speed and direction.</td>
<td>66.67%</td>
</tr>
<tr>
<td>Nguyen Viet Cuong, National Univ of Singapore</td>
<td>SVM with Mel-frequency cepstral coefficients features + split data to 10 segments</td>
<td>48.15%</td>
</tr>
</tbody>
</table>

VII. DISCUSSION

A. Result Differences for Calibrated and Uncalibrated Datasets

Competition results (Table III) indicate that there are substantial biometric accuracy differences between uncalibrated and calibrated datasets presented in this work. It is very important to discuss possible causes for these differences.

1) Calibration Impact

The difference between the ideal and computed behavior scores discussed above for the uncalibrated datasets indicate that the recorded data contains substantial amount of noise and positional error. Similar conclusion can be made after the manual inspection of the data in the post-calibrated form, e.g., it can be seen from Fig. 2 that for Dataset A the eye positional signal is quite far from the stimulus and that signals from each eye are far apart. These outcomes highlight current danger in employing uncalibrated data for the biometric purposes. The
eye movement data collected in the uncalibrated form might 
contain unique noise introduced by a plethora of individual
subject related parameters such as equipment position, head
position, lightning, etc. that might bias identification
outcomes.

2) Impact of Recording Patterns

Histograms presented on Fig. 3 for the uncalibrated datasets
indicate that the majority of the recordings were conducted
with very close temporal proximity of each other, i.e., 95% 
recordings of Dataset A and 83.83% recordings of the Dataset
B were conducted within 60 seconds of each other. We
hypothesize that such recording arrangement made it possible
for the unique subjects’ related noise characteristics discussed
above to be translated to the high identification accuracy.

3) Impact of Other Factors

Other factors that might have contributed to the
identification accuracy differences: 1) uncalibrated datasets
contained the information from both eyes, while calibrated
datasets contained the information from one eye only, 2)
uncalibrated datasets contained more recordings per subject,
which coupled together with the unique noise and recording
patterns produced very high identification accuracy, 3) Dataset
A was available to the participants for the longest period of
time. Substantial amount of re-submissions was done for the
Dataset A and B, providing more opportunities to improve the
results.

For calibrated datasets particular low identification
accuracy result for dataset C can be in part explained by the
fixed vertical stimulus, which reduced the amount of HVS
performance variability that exist between subjects in case of
non-fixed stimulus. This hypothesis is supported by the fact
that for random stimulus presented for the dataset D the
identification accuracy was improved dramatically. We
hypothesize that random stimulus provided an opportunity to
capture subject related HVS variability in the periphery,
where, for example, the impact of nonlineairities present in the
oculomotor plant structure are more pronounced.

B. Recommendations for Creating Future Eye Movement
Databases

We encourage following recommendations when creating
eye movement databases for the biometric testing.

1) Monitoring & Reporting Data Quality

Perform a calibration prior to a recording of each individual
record. Make sure that the average calibration error does not
exceed 1.5 degrees of the visual angle for each eye that is
being recorded. In cases when calibration error exceeds 1.5° it
is important to re-adjust equipment’s settings and re-calibrate
a person to ensure that calibration error below suggested value
is obtained prior to the actual recording. Some additional
information about the impact of calibration error on the
resulting eye movement biometric performance is discussed
here [25]. We suggest that data loss be kept at a
minimum by careful adjustment of the recording setup. We
also hypothesize that data loss up to 10% is reasonable to
ensure high validity of the captured data. However, currently,
there are no detailed studies that measure performance
tradeoffs between the data loss and the corresponding eye
movement biometric accuracy.

In cases when stimulus performance is fixed (i.e.,
predefined pulse-step stimulus) we suggest reporting behavior
scores. Such information would allow assessing: 1) quality of
the recorded data, 2) performance of the eye movement
classification methods in cases when separation of the raw eye
positional signal into fixations and saccades is necessary, 3)
HVS performance. In cases when normal population is
recorded (no HVS pathologies) the behavior scores tend to be
close to their ideal values which are discussed in detail in
[33][35].

2) Temporal Separation of Individual Records

We already discussed the impact of the recordings’
temporal proximity on the accuracy results where temporal
clustering coupled together with the unique noises (e.g.,
Datasets A and B) might lead to the artificially high
identification accuracy. There are no hard guidelines in the
biometric literature that suggest ideal temporal proximity
between the recordings of an individual to validate a new
biometric modality. We can hypothesize that validation
strategies where the times of the recordings for each person
have a uniform temporal spread, e.g., recordings done once
every hour in a day, once per day, once per month, would
allow to minimize biases related to the equipment setup,
stress, fatigue, illness, and drug related effects.

VIII. SUMMARY & FUTURE WORK

Eye movement based biometrics is an emerging field. The
organized competition allowed exploring current state of the
art in terms of the applicable biometric algorithms, outlining
dangers associated with the different approaches related to the
data capture, and providing suggestions for the future creation
of the eye movement biometric databases.

The results of the competition indicate that the information
about the eye movements can be exploited for the biometric
purposes. Current accuracy results can be compared to the
accuracy of the early face recognition systems [30][40],
however we hypothesize that future work will improve current
performance levels. Additional work is required to statistically
measure between and within subject variability of the eye
movement related characteristics and the stability of the eye
movement traits in longitudinal recording where impact of
fatigue, aging, stress, and possible effect of various substances
is present.

As a result of the data processing associated with the
competition it was possible to find current dangers associated
with recording of the uncalibrated eye positional data.
Uncalibrated data makes it very difficult to control the quality
of the data capture, therefore making it possible to record
unique noises coming from the specific adjustments of the
equipment necessary to record a subject. Equipment
 calibration and quality measurement might require additional
time and effort to capture the data, however, it is, currently,
the only way to carefully control what is being recorded and to
ensure the physiological validity of the data. When recording a
subject multiple times we suggest performing and validating a
calibration prior to each recording. Recordings should not be
clustered into a single recording session where a large number of recordings of the same subject are clustered together. In the best case, recordings should be spread over multiple days with equal amount of recordings per day/month/year to avoid the impact of biases introduced by the equipment setup and other factors.

Competitions have the advantage of providing a common ground for comparing the performance of the different biometric methods and techniques. Therefore, we are planning to organize future competitions that would provide such opportunity and will push forward the state of the art of the eye movement-based biometrics.

IX. ACKNOWLEDGMENT

Data collection for Datasets C and D and the competition related work was supported in part by the National Institute of Standards and Technology (NIST) under Grant #60NANB10D213, and Texas State University-San Marcos via internal grants.

REFERENCES


