This is a pre-print

Towards a multi-source fusion approach for eye movement-driven recognition

Ioannis Rigas[a], Evgeny Abdulin[a], Oleg Komogortsev[a]

[a] Department of Computer Science, Texas State University, San Marcos, USA

rigas@txstate.edu, e_a146@txstate.edu, ok11@txstate.edu

Correspondence to:

Dr. Ioannis Rigas

Postdoctoral Research Associate

Department of Computer Science

Texas State University

San Marcos, TX 78666, USA

Phone: (+1)-512-245-0349

Email: rigas@txstate.edu

Abstract

This paper presents a research for the use of multi-source information fusion in the field of eye movement biometrics. In the current state-of-the-art, there are different techniques developed to extract the physical and the behavioral biometric characteristics of the eye movements. In this work, we explore the effects from the multi-source fusion of the heterogeneous information extracted by different biometric algorithms under the presence of diverse visual stimuli. We propose a two-stage fusion approach with the employment of *stimulus-specific* and *algorithm-specific* weights for fusing the information from different matchers based on their identification efficacy. The experimental evaluation performed on a large database of 320 subjects reveals a considerable improvement in biometric recognition accuracy, with minimal equal error rate (EER) of 5.8%, and best case Rank-1 identification rate (Rank-1 IR) of 88.6%. It should be also emphasized that although the concept of *multi-stimulus fusion* is currently evaluated specifically for the eye movement biometrics, it can be adopted by other biometric modalities too, in cases when an exogenous stimulus affects the extraction of the biometric features.

Keywords: eye movement biometrics, multi-stimulus fusion, multi-algorithmic fusion

1. Introduction

The human body provides an invaluable source of distinctive information suitable to be used for the task of biometric recognition [1]. The most well-studied and widely-adopted biometric modalities are the fingerprints, the iris, and the face. Some other explored biometric traits include the palm, the hand geometry, the ears, the nose, and the lips. The analysis of the blood-vessels morphology appears as the main source of biometric features in methods like the vein matching, and the retinal scan. There are also some biometric traits that enfold behavioral characteristics, i.e. traits that are partially connected with the brain activity. Examples of this category involve the speech analysis and voice recognition, the hand-written signature, keystroke dynamics, gait analysis, and the eye movement-driven biometrics. Considering the abundance of the existing biometric modalities and the heterogeneity of the associated features, it may come as no surprise that there is a strong trend in the biometric research towards the investigation and adoption of information fusion techniques.

1.1. Information Fusion in Biometrics

Information fusion can provide numerous benefits in the domain of biometric recognition. The most obvious among them is the expected performance gain in terms of biometric accuracy due to the combination of evidence gathered from multiple cues [2]. Also, the fusion techniques can be employed for the selection and the promotion of the most informative features among a large set of such features [3]. In addition, the combination of different sources of biometric information can open the path for the creation of biometric systems with enhanced robustness against security flaws and spoofing attacks [4].

The fusion of biometric information can be implemented in multiple ways. A common approach is to combine the information coming from different modalities (e.g. fingerprints, face, iris etc.). An early work demonstrating such a multi-modal fusion scheme for fingerprint and face cues was presented in the work of Hong and Jain [5]. Also, one of the first important studies evaluating the information fusion of fingerprint, face, and hand geometry cues was presented by Ross and Jain in [6]. The study presented by Yang et al. [7], investigated the fusion of characteristics that can be extracted exclusively from the hand region, such as the fingerprints, the hand geometry, and the palm-prints.

Analogously, several approaches focused on the fusion of information coming from the face and the head area, given the abundance of distinct characteristics of these specific body regions. In the work of Wang et al. [8], face and iris features were fused in order to combine the virtues of both modalities. The study of Chang et al. [9] involved an appearance-based fusion scheme employing images of the face and the ear. Another category of multi-modal fusion techniques proposed the combination of physical and behavioral biometric cues. Voice and face were among the first combined features [2], [10], whereas other scenarios involved the combination of face and keystroke dynamics [11], and face and gait features [12].

A different type of information fusion in biometrics involves the combination of the data coming from a single biometric modality by applying multi-algorithmic fusion techniques. In the field of fingerprint biometrics there are several examples of information fusion implemented using multiple algorithms in different stages of the recognition process [13], [14], [15]. The work presented by Vatsa et al. [16] employed the iris as the single modality for implementing multi-algorithmic information fusion. Different techniques for performing multi-algorithmic fusion were also evaluated for the face biometrics [17], [18], in an attempt to use the variability of the features of this specific modality. In the work presented by Han and Bhanu [19], a multi-algorithmic scheme was used for the behavioral trait of gait via the analysis of the influence of the external conditions on the gait patterns.

Several multi-instance fusion techniques were developed in an effort to improve the accuracy of the single-modality biometric systems in practical scenarios. The FBI's IAFIS system [20] can capture the fingerprints of all ten fingers and combines the information for producing more accurate results, a technique proven to be particularly robust when operating on large databases. The work presented by Prabhakar and Jain [21] suggested the fusion of the impressions of multiple fingers by employing multiple (four) algorithms, thus creating a scheme for performing both multi-instance and multi-algorithmic fusion. Also, the work presented by Jang et al. [22] proposed a multi-unit fusion approach for the iris biometrics, using the images coming both from the left and from the right eye in order to address the quality issues often occurring when capturing a single instance of the iris.

Irrespectively of the use of a single or multiple modalities, the fusion methods can be also categorized with respect to their involvement in the typical processing levels (modules) followed in the biometric recognition routine [6], i.e. the sensor level, the feature level, the comparison (or matching) score level, and the decision level. Information fusion in the sensor level can be performed by using the data captured by different types of sensors, e.g. optical and capacitance sensors [23]. Fusion in the feature level can be implemented via the direct incorporation of the extracted features into a compact feature representation [9], [24]. However, in several occasions, the nature of the feature vectors prohibits such an operation. The combination of information in the comparison score level is by far the most common strategy for implementing fusion in biometrics [2], [5], [6], [25], [26]. In this case, the universal accessibility of the comparison scores and the minimal influence of the features' heterogeneity act catalytically for the creation of efficient information fusion schemes. Finally, information fusion can be also performed in the classification stage either by using the identification ranking information [27], or by using the decisions regarding the identity or the validity of a verification claim [21, 28].

1.2. Motivation and Contribution

Eye movements are an emerging biometric modality [29], however, the reported performance still lacks the accuracy of the widely adopted modalities, such as the fingerprints and the iris. The existing performance gap can be attributed to the complicated mechanisms involved in the generation of the eye movements, which combine the physical characteristics of the internal eye structure [30], and the behavioral cues related to the brain activity and visual attention [31]. This work presents a multi-source fusion scheme for the combination of eye movement characteristics extracted by different algorithms (multi-algorithmic fusion) under the influence of different visual stimuli (multi-stimulus fusion). Multi-stimulus fusion is a novel concept inspired by the practically proven influence of different visual stimuli on different eye movement-driven biometric algorithms [32], [33], [34]. The theoretical background for performing the multi-stimulus fusion is also supported by several psycho-visual studies, which demonstrate the interrelationships between the visual stimulus and the generated eye movements [35], [36], [37].

The contribution of the current research in the field of eye movement biometrics can be summarized as follows:

1) We introduce the concept of multi-stimulus fusion, i.e. fusion of different instances of the same modality (eye movements) under the influence of different visual stimuli.

2) We propose a hierarchical weighted fusion scheme for the efficient combination of the comparison (matching) scores generated by the different eye movement algorithms (multi-algorithmic fusion) under the influence of diverse visual stimuli. Also, we suggest a weight-training method for the calculation of the fusion weights, which is based on the identification performance of different matchers.

3) We present a comprehensive investigation of the combined effects from the multi-source fusion (multi-stimulus and multi-algorithmic) in the performance of the eye movement-driven biometrics. We provide an extensive analysis regarding the parameters of our model, and demonstrate the achieved performance improvement by using a large database of 320 subjects.

2. Research on Eye Movement Biometrics

The first study on biometric recognition via the eye movements was presented by Kasprowski and Ober [38] a decade ago. It was based on the spectrum analysis of the eye movement signals, and used a randomly 'jumping' point of light as the visual stimulus. The reported False Acceptance Rate (FAR) was 1.36%, and the False Rejection Rate (FRR) was 12.59%. In the work of Bednarik et al. [39], the Fast Fourier Transform (FFT) was used along with the Principal Component Analysis (PCA) for the analysis of the eye movements during the observation of various stimuli (moving cross, images, and text). The achieved Rank-1 IR reached the value of 56%, and the simple form of fusion that was attempted failed to improve the results any further. The work of Kinnunen et al. [40] was inspired from the field of voice recognition, and analyzed the recorded eye movement signals during the observation of complex stimuli (text and video). The reported minimal EER was about 30%. In the work of Komogortsev et al. [32], a model of the internal non-visible structure and functionality of the eye was employed in order to implement the Oculomotor Plant Characteristic (OPC) biometrics. In this case, the visual stimulus was a point of light making horizontal and vertical 'jumps', and the reported Half Total Error Rate (HTER) was 19%. The Complex Eye Movement Behavior (CEM-B) biometrics were introduced by Holland and Komogortsev in [33]. The used visual stimulus consisted of text excerpts, and the fusion of the comparison scores from the individual features led to an EER of 16.5%. An attempt to fuse the information of the OPC and the CEM characteristics was presented by Komogortsev et al. [41], showing a possible performance improvement of 30% over the single methods. In the work of Rigas et al. [42], a graph-based approach was used for comparing the spatial distributions of the eye fixations during the observation of stimulus consisting of human face images. The reported minimal EER was 30%. Face images were also used in the graph-based work of Cantoni et al. [43], where a minimal EER of 25% was reported. In the study of Yoon et al. [44], images of cognition-related dot-patterns were employed as the stimuli in a scheme that used Hidden Markov Models (HMM) to analyze gaze velocity features. The reported Rank-1 IR was in the range of 53%-76%. The recent work of Rigas and Komogortsev [34] suggested a model based on the Fixation Density Maps (FDMs) for representing the eye movements during the observation of dynamically changing stimuli. In the proposed scheme, the information corresponding to the successive time intervals of a video sequence was combined for achieving a minimal EER of 13%.

3. Methodology

3.1. General overview

As already mentioned, the overarching goal of our study was to investigate the effects from the multi-stimulus and multi-algorithmic information fusion in the field of the eye movement biometrics. For this reason, we employed three algorithms originating from different principals, the Oculomotor Plant Characteristic (OPC) biometrics (18-parameter version) [45], the Complex Eye Movement Behavior (CEM-B) biometrics [33], and the Fixation Density Map (FDM) biometrics [34]. The selection of these specific algorithms was decided for the following reasons: a) the features extracted by these algorithms encapsulate information generated by a variety of underlying sources (physical and behavioral), and b) the selected algorithms exhibit stimulus preference, i.e. they can perform more efficiently for specific types of stimulus. Thus, the selected algorithms are suitable for exploring the scenario involving the multi-source information fusion of eye movement-based characteristics. In Fig. 1, we show a graphical overview summarizing the basic properties of the employed eye movement biometric algorithms in terms of the extracted features and the exhibited stimulus preference.

In the current work, we developed a weighted fusion scheme for the combination of the information in the comparison score level. Our decision was mainly driven by the heterogeneity of the features extracted by the employed algorithms (see Section 3.2 for more details), which partially obstructs the application of fusion directly at the feature level. Also, the fusion in the decision level was an unattractive option for the particular scenario where the relative contribution of the different algorithms and visual stimuli needs to be modeled. We should emphasize that although the suggested scheme uses the rank identification performance for implementing information fusion, it should not be conceived as a classical rank level fusion method where the ranking information is directly employed at the decision level. In our method, the ranking information is used to modulate the comparison scores and the fusion is performed in the comparison score level.

In the following section we present a detailed description of the employed eye movement biometric algorithms. Then, in Section 3.3 we present the suggested multi-source weighted fusion scheme.

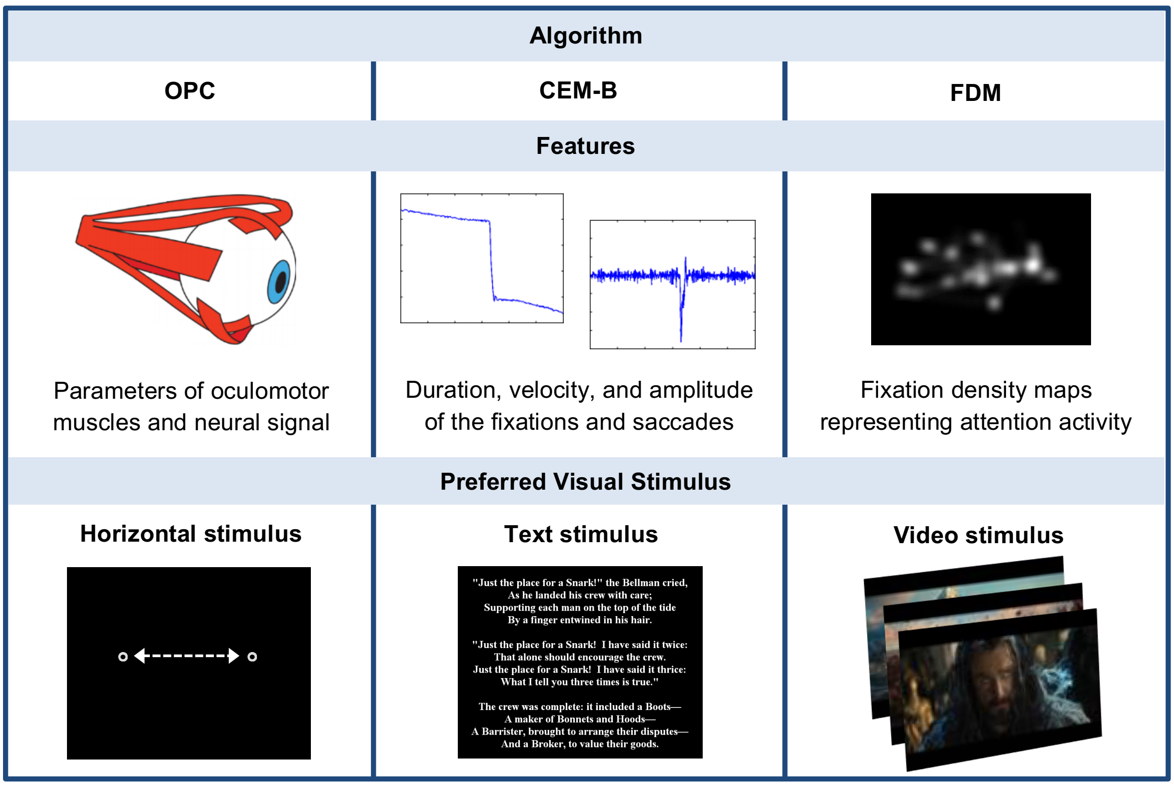

Figure 1. Overview of the basic properties of the employed eye movement biometric algorithms.

3.2. Eye movement biometric algorithms

This section describes the eye movement biometric algorithms used in the current work for the implementation of the multi-source fusion. The presented description aims to unveil the details regarding the features and the comparison modules (matchers) used by each algorithm, and further clarify the rationale behind their selection for the developed weighted fusion scheme.

3.2.1 Oculomotor Plant Characteristic (OPC) biometrics

The algorithm for extracting the OPC features is based on a mathematical model describing the oculomotor system's operation, i.e. the oculomotor plant. The main operation of the algorithm is to simulate the saccadic eye movements and compare them with the actual trajectory made by the real eyes of a user. Thus, the algorithm extracts a number of parameters via the minimization of a cost function during the comparison between the real and the simulated saccadic trajectories. In this work, we adopted the OPC algorithm described in [45], which is supported by a linear homeomorphic 18-parameter model based on the following characteristics: *Series Elasticity (AG/ANT)*, *Length-Tension Relationship (AG/ANT)*, *Force-Velocity Relationship (AG/ANT)*, *Passive Viscosity*, *Tension Slope (AG/ANT)*, *Inertial Mass, Activation Time (AG/ANT)*, *Deactivation Time (AG/ANT)*, *Tension Intercept*, *Neural Pulse (AG/ANT)*, and *Neural Pulse Width*. The abbreviations *AG* and *ANT* denote the parameters corresponding to the agonist and antagonist roles of the extraocular muscles. From each eye movement recording, the OPC biometric template $X^{OPC} \in R^{mxn}$ is formed as a multivariate distribution of *m* samples (one for each saccade) of a *n*-dimensional space (*n* = 18). The comparison module used in the case of the OPC algorithm is the *multivariate Hotelling T^2 test* [46]. In the developed approach, the comparison scores generated by the OPC algorithm (C_{OPC}) are forwarded directly at the input of the multi-source fusion scheme.

3.2.2 Complex Eye Movement Behavior (CEM-B) biometrics

In contrast to the OPC algorithm where the internal structure and functionality of the eye is directly modeled, the algorithm for extracting the CEM-B features [33] analyzes the generated eye movement signals for the extraction of a set of features describing the eye movement dynamics. As the CEM-B algorithm was developed for the extraction of biometric features during complex visual tasks (e.g. text-reading), it can model various properties (physical and cognitive) of the eye fixations and saccades. The extracted features are: *fixation start time*, *fixation duration*, *fixation centroid (horizontal/vertical)*, *saccade start time*, *saccade duration*, *saccade amplitude (horizontal/vertical)*, *saccade mean velocity (horizontal/vertical)*, *saccade peak velocity (horizontal/vertical)*. For each eye movement recording, the CEM-B biometric template $X^{CEMB} = \{x_1(m), x_2(m), \dots, x_n(m)\}$ is formed as an ensemble of n = 12 univariate distributions of m samples (m = number of fixations and saccades). The comparison module used in the case of the CEM-B algorithm is the *Cramer-von Mises two sample test* [47]. In the current approach, the scores from every univariate distribution are summed to form the final comparison scores (C_{CEMB}), which are then forwarded at the input of the multi-source fusion scheme.

3.2.3 Fixation Density Map (FDM) biometrics

The FDM algorithm [34] works by extracting features for the representation of the attention-dependent strategies of the eye movements in the case of dynamically changing stimuli (e.g. video sequences). The extracted features have the form of activation maps, which represent in a probabilistic way the distributions of the fixation point positions. For each eye movement recording, the FDM biometric template $X^{FDM} = \{x_1, x_2, \dots, x_n\}$ is formed as a sequence of *n fixation density maps* x_i (2-D grayscale images) representing the eye movement activity for sequential time intervals. The number of maps (*n*) can be defined dynamically, based on the duration of the visual stimulus presentation, and the selected time interval. The comparison module used in this incarnation of the FDM algorithm was the *similarity metric* [48]. It should be mentioned that although in the original FDM implementation [34] other measures resulted in better performance, during our experiments we verified that the scores extracted with the *similarity metric* are more suitable to be used in the developed multi-source fusion scheme, possibly due to the fact that the *similarity metric* represents an actual metric. In the current approach, the scores from every fixation density map are summed to form the final comparison scores (C_{FDM}), which are then forwarded at the input of the multi-source fusion scheme.

3.3. Multi-source weighted fusion scheme

This section describes the details of the proposed scheme for performing the multi-source information fusion. In Fig. 2, we present a schematic diagram showing the architecture of the developed approach. Let us assume that a user observes different types of visual stimuli (in this example three types) while an eye tracking system captures the performed eye movements. The visual stimuli are presented sequentially to the user, and they can appear in arbitrary order or even have a time gap between them. During the first stage of the developed scheme, multi-stimulus fusion is performed separately for every single biometric algorithm. Initially, each algorithm extracts a number of features, and the corresponding biometric templates are formed. Then, the comparison of the templates is performed, and the calculated comparison scores corresponding to the different visual stimuli are fused using *stimulus-specific* weights. The optimum weight-training method should be apt to quantify effectively—in terms of performance and generalization—the relative contribution of the information deriving from the different stimuli. In this work, we suggest and evaluate a specific weight-training method which is based on the ranking identification performance. During the second stage of the developed scheme, the information fusion is performed via the multi-algorithmic combination of the comparison scores generated during the first stage (multi-stimulus fusion). The multi-algorithmic fusion process quantifies the relative contribution of every algorithm (OPC, CEM-B, FDM) via the use of *algorithm-specific* weights. Prior to their final combination, the comparison scores need to be normalized with the use of an appropriate normalization function. We should note that the separation of the multi-source fusion procedure in two distinct stages allows for the investigation of the relative importance of the two different types of fusion (multi-stimulus vs. multi-algorithmic). Also, the suggested scheme provides flexibility and robustness since it allows for the separate training of the weights used for the two types of fusion, and if required, it permits the application of different normalization functions in the two stages.

The developed multi-source fusion scheme can be mathematically described in the form of the general equation:

$$C_{fused} = \sum_{i=1}^{N} w_{a_i} \cdot f_n(S_i),\ S_i = \sum_{j=1}^{M} w_{s_i^j}\, C_i^j \quad (1)$$

In this formula, we denote with i the index of the biometric algorithm and with j the index of the stimulus type. Thus, the term C_i^j represents the summed comparison scores extracted from a specific algorithm i during the presentation of a specific stimulus j. The *stimulus-specific* and *algorithm-specific* weights are denoted with $w_{s_i^j}$ and w_{a_i} respectively. Finally, we denote with $f_n(\cdot)$ the normalization function used prior to the multi-algorithmic combination. The operation of the block diagram shown in Fig. 2 can be derived by Eq. (1) using the following parameters: N = 3 (OPC, CEM-B, FDM) and M = 3 ('jumping' point, text, and video). It should be noted that during our experiments we used four types of visual stimulus, since the 'jumping' point stimulus consisted of two sub-cases, the horizontally 'jumping' point and randomly 'jumping' point.

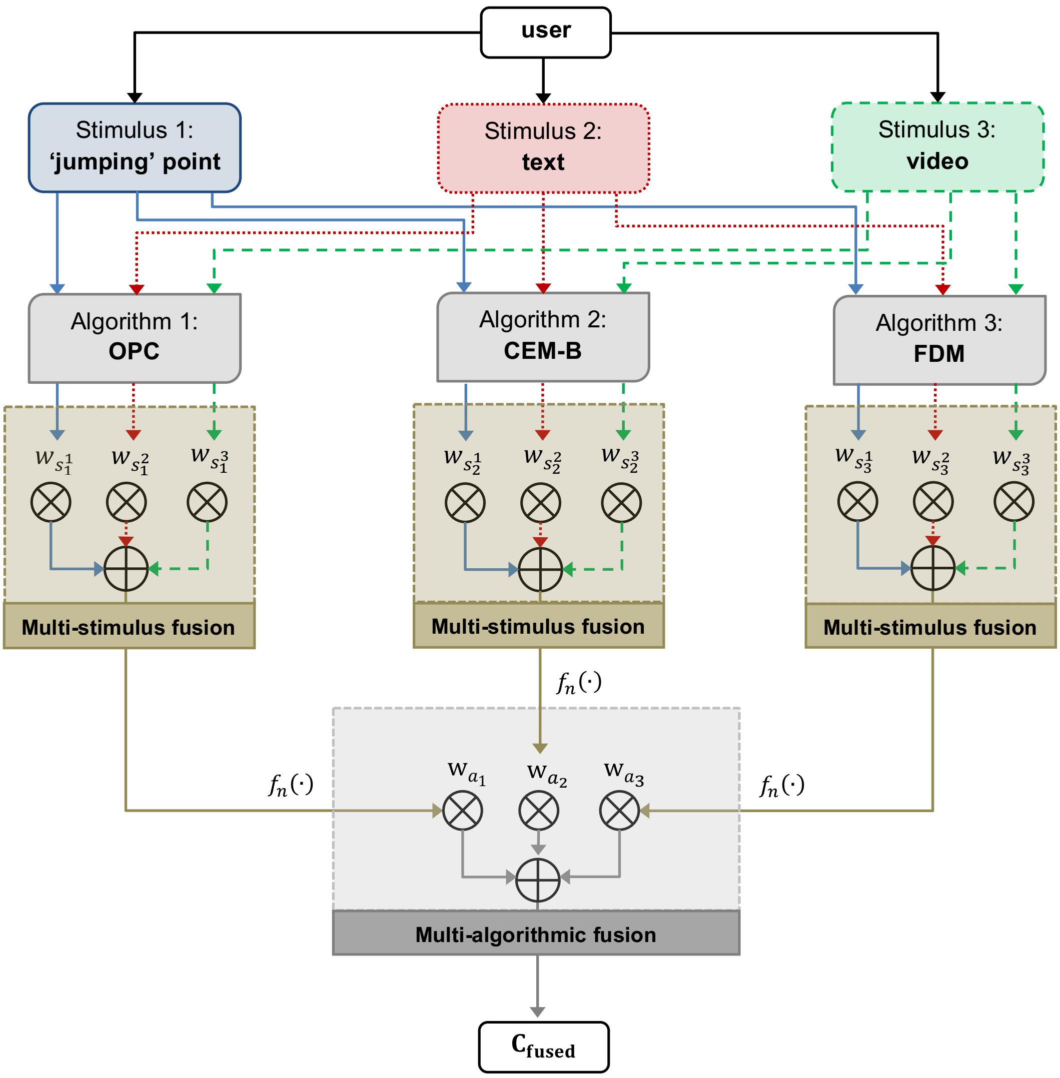

Figure 2. Diagram of the suggested scheme for performing multi-source fusion based on eye movement cues.

3.3.1 Normalization of the comparison scores

Typically, the normalization procedure is required because the comparison modules employed by the different algorithms usually result in the generation of scores which are dissimilar in their distribution and numerical range. In the past, several methods of score normalization were proposed [49], addressing different issues involved in the fusion process. In this work, we evaluated two normalization techniques for our scheme: the *Max-Min* normalization technique (*MM*) and the *Z-score* normalization technique (*ZS*). In what follows, let us denote the set of K comparison scores which need to be normalized as: $C \rightarrow \{c_k\}, k = 1,2, \dots, K$, and the resulting set of normalized scores as: $N \rightarrow \{n_k\}, k = 1,2, \dots, K$.

3.3.1.1 Max-Min normalization technique (MM)

The *Max-Min* technique provides a simple and efficient approach for the normalization of the comparison scores. In this technique the comparison scores are normalized based on the maximum and the minimum values appearing in a set of scores. This approach has two important advantages: on the first hand, the scores are transformed into a fixed common range [0, 1]. On the other hand, the original form of the distribution of scores is retained. The *Max-Min* normalization technique can be implemented using the following formula:

$$n_k = \frac{c_k - \min_k C}{\max_k C - \min_k C} \tag{2}$$

3.3.1.2 Z-score normalization technique (ZS)

The normalization using the *Z-score* technique is performed via the calculation of the arithmetic mean and the standard deviation of a set of scores. The resulting distribution of scores has a mean of zero and a standard deviation of one. A disadvantage of the *Z-score* normalization is that it does not guarantee a common numerical range for the normalized scores. The *Z-score* normalization technique can be implemented using the following formula:

$$n_k = \frac{c_k - \text{mean}\,(C)}{mean(C) - std(C)} \tag{3}$$

It should be noted that although both the *Max-Min* and the *Z-score* normalization techniques can be sensitive to the presence of outliers, their performance in the scope of the proposed multi-source fusion approach was found to be satisfactory in all cases. Furthermore, during our experiments we also evaluated the *Hyperbolic Tangent Estimators* normalization technique [49], a method which presents robustness in the presence of outliers. The particular technique performed with acceptable rates for the verification scenario but resulted in poor performance in the identification scenario. Thus, it was considered as unsuitable to be also included in the analysis of the developed multi-source fusion scheme.

3.3.2 Computation of the multi-source fusion weights

In this section we describe the procedure followed for the calculation of the multi-source fusion weights based on the rank identification performance, with a special focus on the case of the Rank-1 identification performance. Furthermore, we briefly present a more traditional weight-training procedure based on the verification performance (equal error rate - EER), a method originally proposed in [50]. During the evaluation process, the three weight-training methods (Rank, Rank-1, and EER) are compared and their special characteristics are discussed in details.

3.3.2.1 Weight-training method based on the rank identification performance

Let us denote with R the full ranked list formed using the comparison scores computed for a probe item i with all the reference items—all items refer to a training dataset. Also, we denote with r_i the rank of the first reference item in the list corresponding to the same identity with item i. In this case, we can calculate the corresponding rank weight $w(i)$ for each one of the K probe items as:

$$w(i) = 1 - \frac{r_i - 1}{|R|}, \quad i = 1, 2, \dots, K \qquad (4)$$

In the case of r_i=1 (item ranked first) the weight equals to one, whereas for $1 < r_i < |R|$ the weight value becomes successively lower as it approximates zero. By calculating the weights for all the K probe items, we can compute the total rank weight w'_m for a specific matcher (m) as:

$$w'_m = \frac{\sum_{i=1}^{K} w(i)}{K} \qquad (5)$$

It should be noted, that in our case the term ‘matcher’ is used to denote both the modules that extract scores from different stimuli (*multi-stimulus fusion*), and from different algorithms (*multi-algorithmic fusion*). The final rank identification-trained weights are calculated by normalizing all the weights to sum to unity:

$$w_m^{Rank} = \frac{w'_m}{\sum_{m=1}^{M} w'_m} \tag{6}$$

Now, let us consider the special case of using the Rank-1 identification rate for training the weights. This case may be considered as a sub-category of the previous problem: here, instead of a full ranked list containing ranks for all the reference items of the dataset, an individual probe item i can have either a first rank match $(r_i = 1, w(i) = 1)$ or not $(w(i) = 0)$. The total weight can be calculated by averaging the weights for all the probe items in the training dataset, and the final Rank-1 based weight w_m^{Rank1} of a matcher can be calculated again using Eq. (6).

3.3.2.2 Weight-training method based on the verification performance

The weighting of the matchers based on the use of the equal error rate (EER) performance was originally proposed in [50]. In this particular case, the fusion weights are calculated based on the behavior of different matchers in the verification scenario. Using a training dataset containing the comparison scores coming from different matchers, the corresponding Receiver Operating Characteristic (ROC) curves can be constructed, and consequently, the EER performances e_m can be calculated and employed as the training criterion for the weights. The weights are inversely analogous to the EER performance, and may be calculated using the following formula to ensure that they are normalized to unity:

$$w_m^{EER} = \frac{1/\sum_{m=1}^{M} \frac{1}{e_m}}{e_m} \tag{7}$$

3.3.3 Weights transformation

After the calculation of the weights performed by either of the above described methods, we opted to implement a transformation procedure aiming at the optimization (fine-tuning) of the weight values. The specific procedure can be performed via the application of a single-parameter linear range transformation, described by the following equation:

$$w_m^{final} = \frac{w_m - \min_m(W)}{\max_m(W) - \min_m(W)} \cdot \left(1 - w_{opt}\right) + w_{opt} \tag{8}$$

Here, w_m is the weight of a specific matcher, W is the set of weights from all the matchers, and w_{opt} is the optimization parameter that can be automatically trained by monitoring the escalation of the recognition rates in a development (training) dataset. The exact details regarding the training process and the final global value selected for the optimization parameter are described in Section 5.2.

4. Experimental procedure

4.1. Visual stimuli

The eye movement recordings were performed using the following categories of visual stimuli: two types of *'jumping' point stimulus* (HOR and RAN)*, text stimulus* (TEX)*, and video stimulus* (VID).
In the case of the *'jumping' point stimulus* a white circular point of light with a black center was making 'jumps' in the black background of a computer screen, jumping from one position to another at predefined time intervals of 1 second. The participants were instructed to follow the point with their eyes, forcing thus the execution of eye saccades. There were two separate experiments involving the *'jumping' point stimulus*: the horizontally 'jumping' point (HOR), inducing horizontal saccades, and the randomly 'jumping' point (RAN), inducing random oblique saccades. The total duration of each experimental trial was 1 minute and 40 seconds.
In the case of the *text stimulus* (TEX) a number of text excerpts were presented in a computer screen, and the participants were instructed to freely read them. The used excerpts were from the poem of Lewis Carroll "The Hunting of the Snark". The specific poem was chosen due to its specific writing style, which encourages the observer to actively process the text while reading it. The total time given to the participants to read the text excerpts in each experimental trial was 1 minute.
In the case of the *video stimulus* (VID) a segment from a movie trailer was presented on a computer screen, and the participants were instructed to freely observe the video. The chosen video segment was from the official trailer of the Hollywood film "Hobbit 2: The Desolation of Smaug (2013)". The specific trailer was used due to the diversity of its content, which contains both dynamic action scenes and static parts with emotional content. The total duration of each experimental trial was 1 minute.

4.2. Participants

The experiments for the collection of the eye movement recordings were performed with the participation of 320 subjects (170 males/150 females), ages 18-46, (M = 22, STD = 4.23). Texas State

University's institutional review board approved the study, and the participants provided informed consent. Every subject participated in two recordings sessions. The time interval separating the two recording sessions was approximately 20 minutes. In every session, the four used types of visual stimulus were presented on a computer screen while the eye movements of the participant were recorded. This led to the formation of a database of 2560 unique eye movement recordings. Between the experimental trials for each visual stimulus the subjects performed various eye movement tasks and had short periods of rest to mitigate eye fatigue.

4.3. Apparatus

The eye movement recordings were performed using an EyeLink 1000 eye tracker [51], with a sampling frequency of 1000 Hz. The device has a vendor reported spatial accuracy of 0.5° and a spatial resolution of 0.01° RMS. The capturing device was positioned at a distance of 550 millimeters from the computer screen where the stimulus was presented. The size of the computer screen was 474 x 297 millimeters and the resolution 1680 x 1050 pixels. The heads of subjects were comfortably stabilized with the use of a chin-rest with a forehead in order to ensure the high quality of the recorded data. The quality of the capturing procedure was evaluated with the experimental measurement of the *calibration accuracy* and the *recording validity* (i.e. number of samples indicated by the device as tracked). The recorded datasets were captured with a measured average *calibration accuracy* of 0.49° (STD = 0.17°), and an average *recording validity* of 96.77% (STD = 4.96%).

4.4. Datasets partitioning

The recordings for the different visual stimuli were used to form four separate datasets denoted with the corresponding abbreviations: HOR, RAN, TEX, and VID. We performed 20 random splits in order to partition each dataset in development and evaluation sets. In every split, each dataset was partitioned in two halves. All the data from half of the subjects (160 subjects) were employed for the development set, and used for training the multi-source fusion weights. All the data from the other half of the subjects (160 subjects) were employed for the evaluation set, and used for the evaluation procedure. It should be emphasized that the partitioning of the data in development and evaluation sets was done with no overlap of the used subjects in order to ensure that the evaluation procedure will not be affected by any kind of overfitting effects. All the experimental results presented in the following sections were extracted by taking the average over the above-mentioned 20 random splits.

5. Results Evaluation

5.1. Performance evaluation metrics

Rank-1 Identification Rate (Rank-1 IR): during the identification scenario, the biometric system aims to detect the real identity of a user by comparing the current biometric sample with the templates stored in the database. The most popular metric for the evaluation of the identification accuracy is the Rank-1 Identification Rate, defined as the ratio of the testing samples that were assigned to the correct identity divided to the total number of the testing samples of the dataset.

Equal Error Rate (EER): during the verification scenario, the biometric system aims to check the validity of a claim of a user that his/her biometric template belongs in the database. A user whose the template belongs in the database is called a genuine user, whereas a user that does not belong in the database is called an impostor. The correct acceptance of a genuine user from the system raises the Genuine Acceptance Rate (GAR). Inversely, the false acceptance of an impostor from the system raises the False Acceptance Rate (FAR). Finally, the false rejection of a genuine user from the system raises the False Rejection Rate (FRR). A Receiver Operating Characteristic (ROC) curve can be constructed by changing the acceptance threshold and calculate the respective GAR and FAR. The EER can be computed as the point of the ROC curve where the FAR equals the FRR (FRR = 100% - GAR). In this work, we used the vertical averaging technique described in [52] for averaging the ROC curves constructed from the 20 random splits.

GAR at 0.1% FAR: this measure can be used to complementarily assess the verification accuracy of a biometric system. The GAR at 0.1% FAR expresses the verification performance of a biometric system in the region of the low FAR values, which is usually of particular importance. We decided to use this additional measure in order to perform a more detailed comparison of the tested weight-training methods during the task of multi-source information fusion (see Section 5.5).

5.2. Training of the weight optimization parameter (w_{opt})

In Section 3.3.3, we described the transformation step performed for the optimization of the weight values. In order to calculate the exact value for the optimization parameter w_{opt} we performed a training procedure using the development datasets. For each of the tested weight-training methods (WM^{Rank1}, WM^{Rank}, WM^{EER}) we scanned the range of the allowed w_{opt} values [0, 1], and calculated the resulting identification and verification performances (in terms of the achieved Rank-1 IR and EER). In Fig. 3, we show the effects of varying the value of the optimization parameter w_{opt} for each separate weight-training method. Each diagram is shown in a double-vertical axis mode so that the co-variation of the Rank-1 IR and EER performances can be inspected in tandem. Only the values in the range [0.05, 0.75] are shown, since out of these bounds the performance deteriorates considerably.

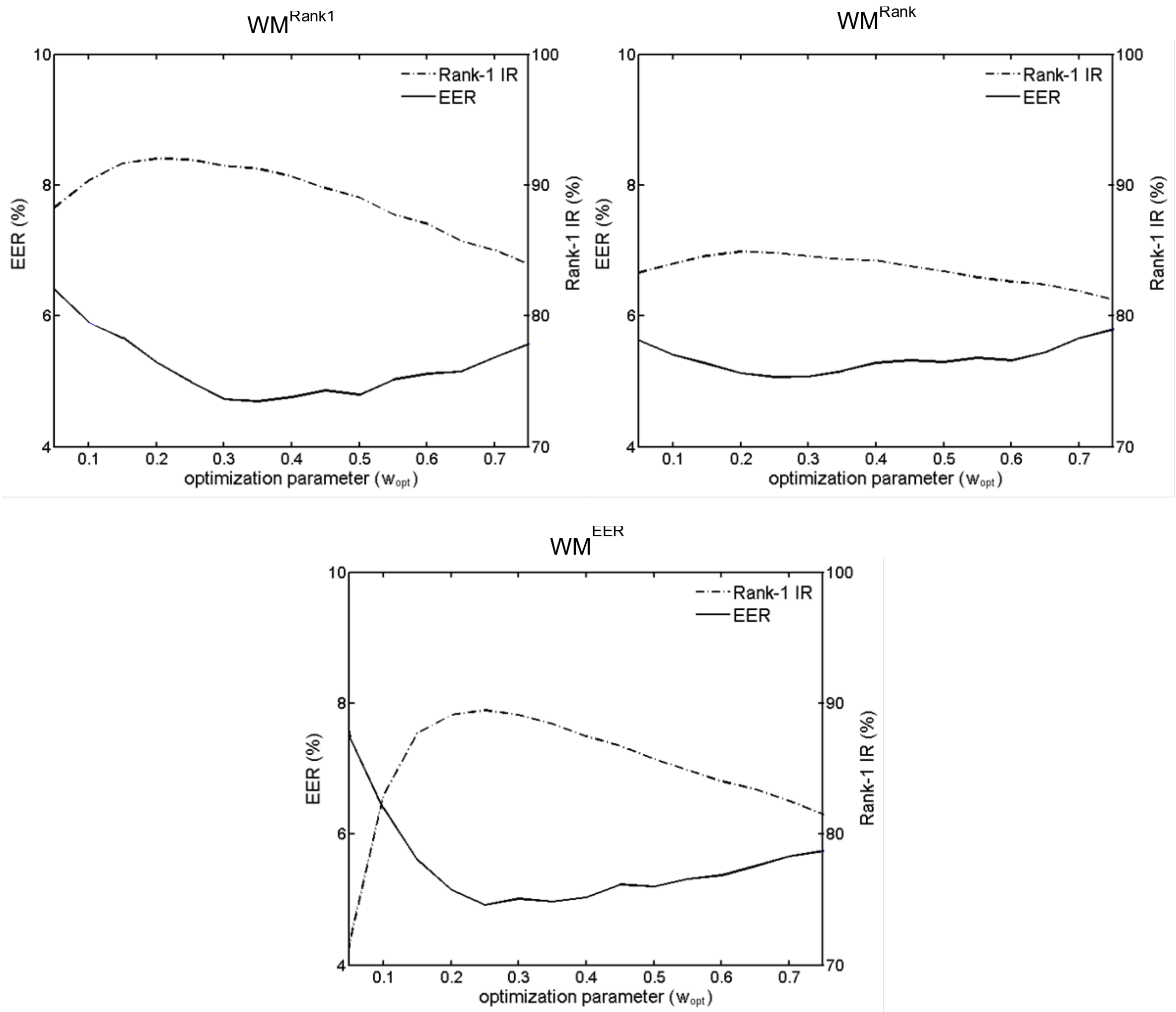

Figure 3. Performance curves demonstrating the dynamics of the joint training procedure used for the selection of the global value for parameter w_{opt}.

As we can observe, the exact points (w_{opt} values) where the EER minimization and the Rank-1 IR maximization occur can be slightly different. Thus, in order to select a common value for the optimization parameter (w_{opt}) that can be used for all the weight-training methods and for both recognition scenarios (identification and verification) we adopted the following joint optimization rules:

$$w_{opt} = mean\left(w_{opt}^{IR}, w_{opt}^{EER}\right) \quad (9)$$

$$w_{opt}^{IR} = mean\left(\underset{w}{argmax}\, IR(WM^{Rank1}), \underset{w}{argmax}\, IR(WM^{Rank}), \underset{w}{argmax}\, IR(WM^{EER})\right) \quad (10)$$

$$w_{opt}^{EER} = mean\left(\underset{w}{argmin}\, EER(WM^{Rank1}), \underset{w}{argmin}\, EER(WM^{Rank}), \underset{w}{argmin}\, EER(WM^{EER})\right) \quad (11)$$

Using the development set and the Eq. (9-11) we calculated the globally optimal value $w_{opt} = 0.26$, which was then routinely used throughout our experiments.

5.3. Analysis of the multi-source fusion weights

In this section, we present a detailed analysis for the multi-source fusion weights trained with the three tested weight-training methods. This analysis demonstrates the efficacy of the used algorithms in modeling the fusion weights, and additionally it provides further insights for our motivation to extract and combine *stimulus-specific* and *algorithm-specific* weights during the fusion process.

In Fig. 4, we show a comparison of the trained *stimulus-specific* weights $\left(w_s^{Rank1}, w_s^{Rank}, w_s^{EER}\right)$ for the three weight-training methods. These bar diagrams are created by averaging the calculated weight values using the data from the development sets (20 random splits). We also show the corresponding error bars with the error margins in 95% confidence intervals. A close inspection of the trained weights provides the first practical evidence regarding the stimulus-preference exhibited by the different eye movement biometric algorithms. For the OPC algorithm, the weight contribution of the *horizontal 'jumping' point stimulus* (HOR) clearly predominates compared to the other types of visual stimuli. The stimulus with the least contribution in this case is the *video stimulus* (VID). For the CEM-B algorithm, the calculated weights reveal that the *text stimulus* (TEX) is the preferred type of stimulus since it presents the larger value across all weight-training methods. Finally, for the FDM algorithm, the larger weight values are assigned to the *video stimulus* (VID).

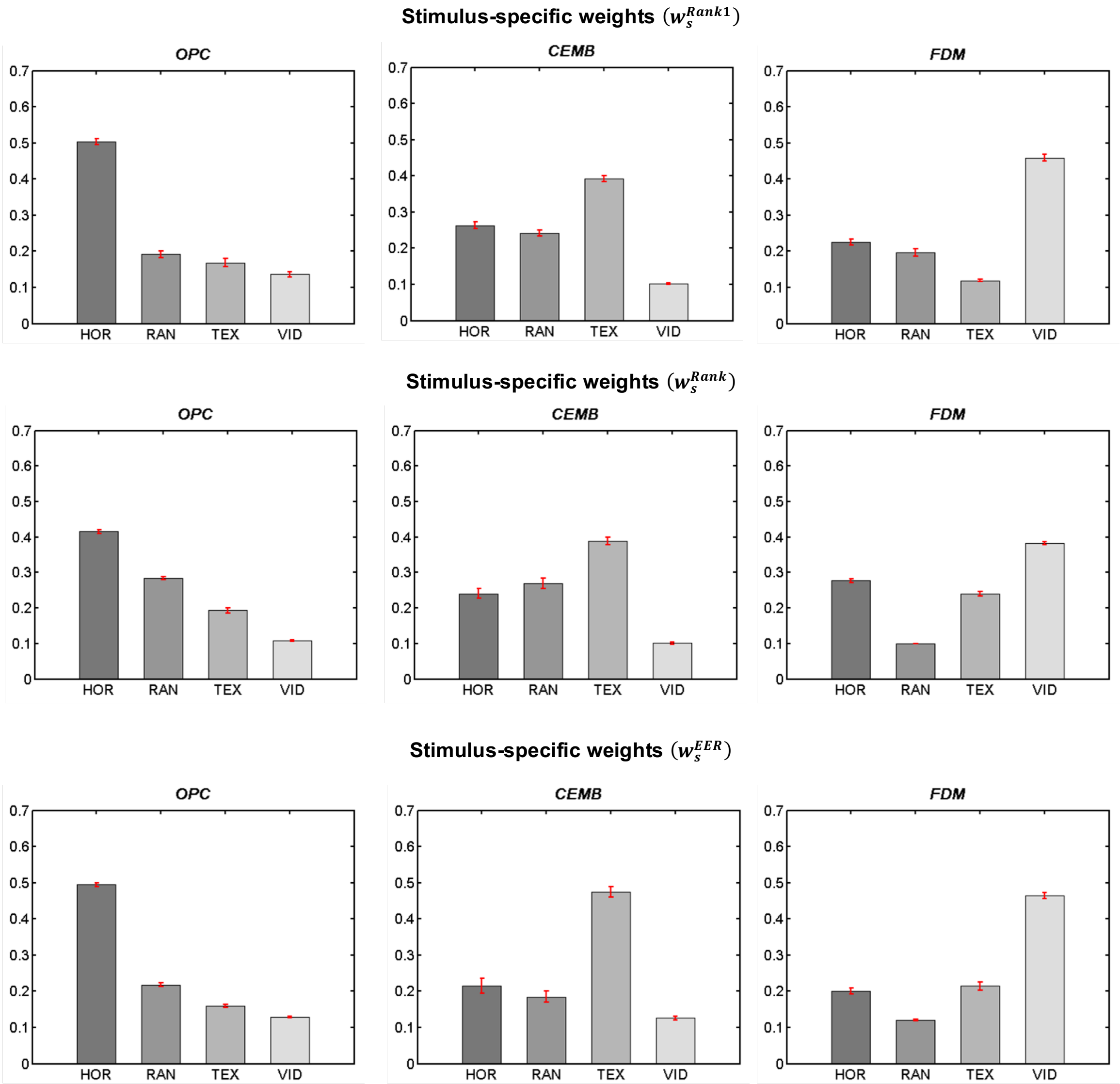

Figure 4. Diagrams of the trained weights $(w_s^{Rank1}, w_s^{Rank}, w_s^{EER})$ for the three tested weight-training methods, in the case of the multi-stimulus fusion. The error bars correspond to 95% confidence intervals.

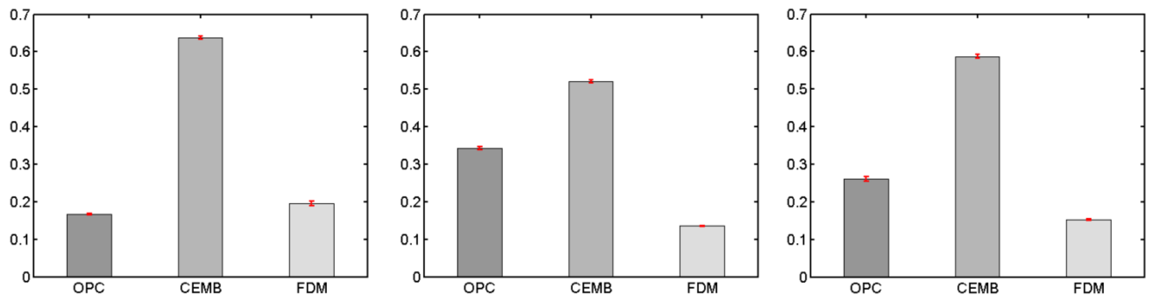

Figure 5. Diagrams of the trained weights $(w_a^{Rank1}, w_a^{Rank}, w_a^{EER})$ for the three tested weight-training methods, in the case of the multi-algorithmic fusion. The error bars correspond to 95% confidence intervals.

All the three tested weight-training methods can model consistently the stimulus preference characteristics of each of the biometric algorithms, thus confirming the behavior anticipated from the theoretical analysis. An important observation is that the contribution from the other types of stimuli can be also relevant, and that the relative significance of this contribution can vary based on the selected biometric algorithm. The calculated error margins are sufficiently low (< 0.021), revealing thus the high stability of the weight-training procedure in all cases. A one-way ANOVA for the error margins across the weight-training methods (for all stimuli and all algorithms) showed no significant main effect $F(2, 33) = 0.76$, $p = 0.48$, supporting the equivalent behavior of the weight-training methods in terms of stability.

In Fig. 5, we present the corresponding diagrams for the weights trained during the multi-algorithmic fusion stage $\left(w_a^{Rank1}, w_a^{Rank}, w_a^{EER}\right)$, which is applied immediately after the multi-stimulus fusion stage. The comparative overview of the calculated weights for the three used biometric algorithms shows the strong dominance of the CEM-B algorithm across all weight-training methods. As we show in the next section, the dominance of the CEM-B algorithm practically reflects the large performance improvement for this specific algorithm during the first stage of fusion, i.e. the multi-stimulus fusion. We should emphasize, though, that the weights for the other two algorithms are not negligible, and they can practically contribute to the further improvement of the biometric recognition performance. In terms of stability, the behavior of the three weight-training methods is even better than previously, with the error margins in 95% confidence interval being lower than 0.007 in all cases.

5.4. Single algorithm multi-stimulus fusion performance

In this section, we evaluate the effects of the multi-stimulus fusion (first stage of fusion) in the performance of each of the employed biometric algorithms. Table 1 shows the baseline Rank-1 IR performances achieved by each of the biometric algorithms for every type of visual stimulus separately, and also, the respective rates obtained after the application of the multi-stimulus fusion (M-ST) with the use of the three tested weight-training methods. A first observation is that the baseline performances seem to confirm the already discussed stimulus preference exhibited by the different biometric algorithms. For the OPC algorithm, the top baseline Rank-1 IR is 21.50% and occurs for the HOR stimulus. The CEM-B algorithm presents a top baseline Rank-1 IR of 47.45% for the TEX stimulus, and the FDM algorithm performs with a top baseline Rank-1 IR of 27.12% for the VID stimulus. After the application of the multi-stimulus fusion the identification rates improve considerably, reaching the best case values of 28.40% for the OPC algorithm, 82.03% for the CEM-B algorithm, and 30.69% for the FDM algorithm. A close inspection of the rates achieved from the three tested weight-training methods (M-ST^{Rank1}, M-ST^{Rank}, M-ST^{EER}) portrays the differentiations in their performances. A one-way ANOVA (using values from 20 random splits) revealed a main significant effect for Rank-1 IR across the evaluated weight-training methods in all cases, with $F(2, 57) = 5.68$, $p < 0.01$ for the OPC algorithm, $F(2, 57) = 11.87$, $p < 0.001$ for the CEM-B algorithm, and $F(2, 57) = 56.32$, $p < 0.001$ for the FDM algorithm.

Table 1. The Rank-1 IR performances in the case of the multi-stimulus (M-ST) fusion for each single biometric algorithm.

Rank-1 Identification Rate (STD) %							
Algorithm	*Single Stimulus Baselines*				*Multi-Stimulus Fusion*		
	HOR	**RAN**	**TEX**	**VID**	**M-ST^{Rank1}**	**M-ST^{Rank}**	**M-ST^{EER}**
OPC	**21.50** (3.26)	7.17 (1.18)	7.39 (1.54)	5.11 (0.97)	28.19 (2.26)	26.05 (2.42)	**28.40** (2.62)
CEM-B	33.83 (2.77)	32.44 (3.15)	**47.45** (2.53)	16.86 (2.21)	**82.03** (2.01)	81.42 (2.40)	78.55 (2.77)
FDM	10.41 (1.56)	7.98 (1.27)	3.69 (1.28)	**27.12** (2.64)	**30.69 (2.47)**	19.45 (3.29)	23.83 (4.15)

In Table 2 we show the respective performance results for the verification scenario. In this case, the baseline EER values are 14.43% for the OPC algorithm and the HOR stimulus, 15.01% for the CEM-B algorithm and the TEX stimulus, and 26.93% for the FDM algorithm and the VID stimulus. As for the case of the identification scenario, the application of the multi-stimulus fusion leads to a generalized improvement of the verification rates, with the calculated best case values for the EER reaching 13.72% for the OPC algorithm, 7.28% for the CEM-B algorithm, and 22.97% for the FDM algorithm.

Table 2. The EER performances in the case of the multi-stimulus (M-ST) fusion for each single biometric algorithm.

Equal Error Rate (STD) %							
Algorithm	*Single Stimulus Baselines*				*Multi-stimulus Fusion*		
	HOR	**RAN**	**TEX**	**VID**	**M-ST^{Rank1}**	**M-ST^{Rank}**	**M-ST^{EER}**
OPC	**14.43** (0.73)	21.54 (0.85)	25.09 (1.06)	28.09 (1.36)	13.86 (0.85)	13.83 (0.90)	**13.72** (0.86)
CEM-B	18.39 (1.24)	20.21 (1.58)	**15.01** (1.11)	22.78 (1.31)	7.50 (1.12)	7.92 (1.27)	**7.28** (1.05)
FDM	35.03 (1.35)	44.07 (1.10)	35.28 (1.53)	**26.93** (1.21)	**22.97** (1.07)	24.12 (1.40)	23.34 (1.19)

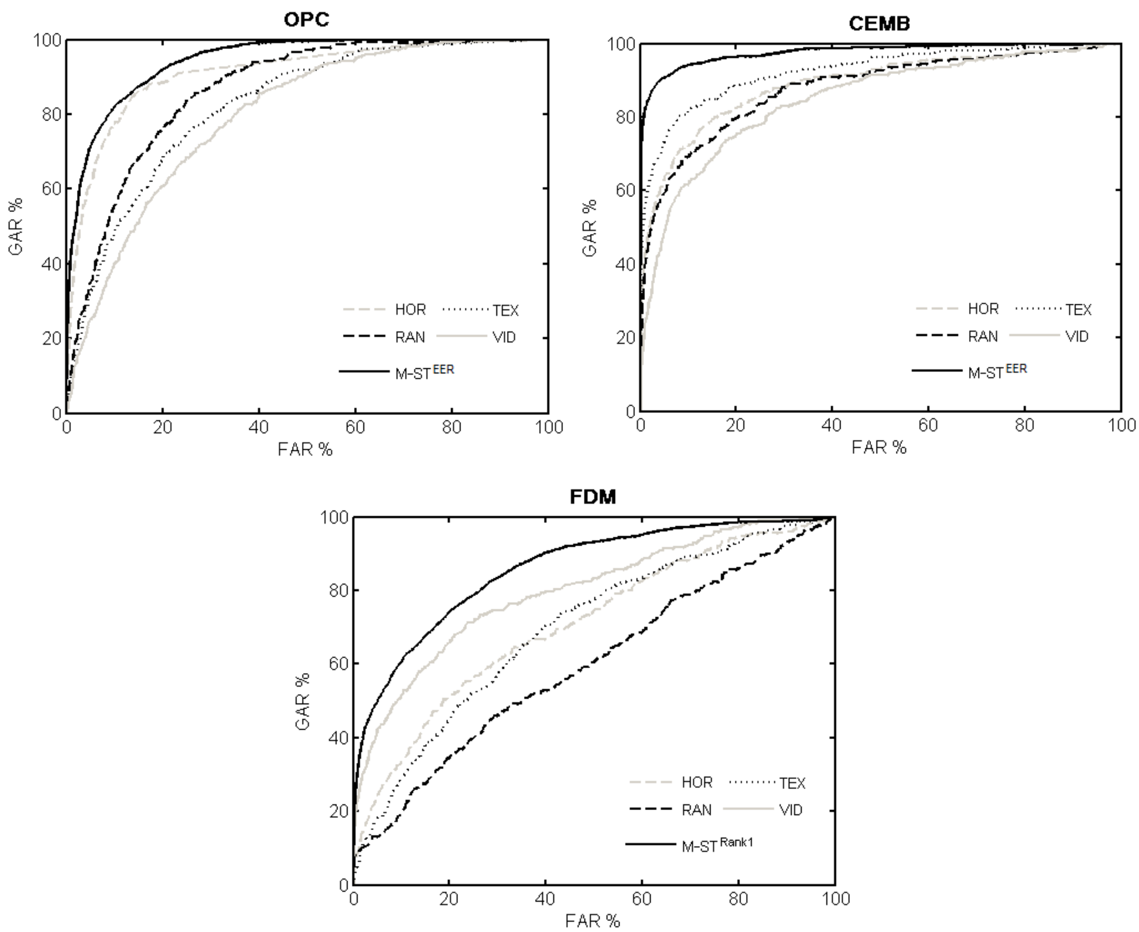

Figure 6. The constructed ROC curves for each single biometric algorithm before and after the application of the multi-stimulus (M-ST) fusion.

The largest improvement in the verification performance was again achieved by the CEM-B algorithm. However, in the case of the EER the differences among the three tested weight-training methods are less noticeable. A one-way ANOVA (using values from 20 random splits) revealed no significant main effect for the EER values across the weight-training methods, with $F(2, 57) = 0.15$, $p = 0.87$ for the OPC algorithm, $F(2, 57) = 1.61$, $p = 0.21$ for the CEM-B algorithm, and $F(2, 57) = 4.54$, $p = 0.01$ for the FDM algorithm. In Fig. 6, we present the constructed ROC curve clusters, which exhibit the overall performance before and after the multi-stimulus fusion for each single biometric algorithm (OPC, CEM-B, FDM). In each case, we show the baseline ROC curves for every single visual stimulus (HOR, RAN, TEX, VID), and the resulting ROC curve after the application of the multi-stimulus fusion using the best performing weight-training method in each case.

5.5. Multiple algorithm multi-stimulus fusion performance

In this section, we present the achieved performances for the case of the multi-source (M-SRC) fusion, i.e. application of both stages of fusion—multi-stimulus followed by multi-algorithmic fusion. In Table 3, we show the Rank-1 IR values obtained by the three weight-training methods and the two tested normalization schemes. For comparison reasons, we also show the achieved rates using two other fusion approaches: the first one is the simple mean (SM) fusion (equivalent to the sum rule fusion), and the second is a method following a different rationale (a classification based approach) with the use of the Random Forests (RF) fusion algorithm [53]. Our current experiments were implemented using the regression Random Forests algorithm with the number of trees set to 100. It should be noted that during our preliminary experiments we also tested other fusion approaches, such as the product rule, the maximum rule, and the minimum rule. These approaches did not succeed on producing any competitive rates, and as a result they were not included in our analysis. An inspection of the values in Table 3 reveals that, in all cases, the weighted fusion methods outperform both the SM fusion method and the RF fusion method. The multi-source fusion using the weight-training method based on the Rank-1 identification performance (M-SRC^{Rank1}) achieves the top Rank-1 IR of 88.62%, whereas the other two weight-training methods (M-SRC^{Rank}, M-SRC^{EER}) achieve competitive but lower rates of 81.02% and 84.36% respectively. The Random Forests (RF) fusion method performs with a Rank-1 IR of 80.48%, whereas the Simple Mean (SM) fusion method achieves a lower rate of 76.83%. A one-way ANOVA (using values from the 20 random splits) for Rank-1 IR across the three tested weight-training methods (M-SRC^{Rank1}, M-SRC^{Rank}, M-SRC^{EER}) verifies that the exhibited differences in performance are statistically significant, both for the Max-Min (MM) normalization scheme $F(2, 57) = 107.89$, $p < 0.001$, and for the Z-Score (ZS) normalization scheme $F(2, 57) = 84.34$, $p < 0.001$. However, the one-way ANOVA for Rank-1 IR across the normalization schemes (using 20 random splits and all weight-training methods) revealed no significant main effect $F(1, 118) = 0.16$, $p = 0.69$.

In Table 4, we present the corresponding EER performances for the verification scenario. In this case, the M-SRC^{Rank} method leads to the optimum rates, with the minimal EER of 5.83%. The M-SRC^{EER} method presents an EER of 5.88%, and the M-SRC^{Rank1} scheme achieves an EER of 6.03%. Furthermore, the corresponding EER values for the Random Forests (RF) and the Simple Mean (SM) methods reach the comparable levels of 6.03% and 6.57% respectively. In contrast to the case of the Rank-1 IR, in this case the variation in performance for the three tested weight-training methods (M-SRC^{Rank1}, M-SRC^{Rank}, M-SRC^{EER}) is not statistically significant. This can be verified by the results of the one-way ANOVA (using values from 20 random splits) for the EER values across the weight-training methods, revealing no statistical significant effect both for the Max-Min (MM) normalization scheme $F(2, 57) = 0.27$, $p = 0.76$, and for the Z-Score (ZS) normalization scheme $F(2, 57) = 0.29$, $p = 0.75$. As for the case of the Rank-1 IR, the selection of a specific normalization scheme seems to have a low impact on the EER performance, since the one-way ANOVA for the EER values across the normalization schemes revealed no statistically significant effect $F(1, 118) = 0.07$, $p = 0.79$.

Table 3. The Rank-1 IR performances in the case of the multi-source (M-SRC) fusion.

Rank-1 Identification Rate (STD) %					
Normalization	*Multi-source (multi-stimulus and multi-algorithmic) Fusion*				
	M-SRC^{Rank1}	**M-SRC^{Rank}**	**M-SRC^{EER}**	**RF**	**SM**
MM	**88.62** (1.43)	80.62 (2.05)	84.36 (1.63)	80.25 (2.48)	72.03 (2.83)
ZS	**88.19** (1.50)	81.02 (2.12)	83.62 (1.61)	80.48 (2.95)	76.83 (2.06)

Table 4. The EER performances in the case of the multi-source (M-SRC) fusion.

Equal Error Rate (STD) %					
Normalization	*Multi-source (multi-stimulus and multi-algorithmic) Fusion*				
	M-SRC^{Rank1}	**M-SRC^{Rank}**	**M-SRC^{EER}**	**RF**	**SM**
MM	6.03 (0.86)	**5.83** (0.85)	5.92 (0.96)	6.09 (0.65)	6.79 (0.84)
ZS	6.08 (0.86)	5.94 (0.86)	**5.88** (0.88)	6.03 (0.85)	6.57 (0.71)

In order to provide a more comprehensive analysis of the performance differences among the three tested weight-training fusion methods, we opted to use the complementary performance measure of GAR at 0.1% FAR, for inspecting their behavior in the critical area of the low FAR values. Table 5 shows the calculated values for the GAR at 0.1% FAR for the three weight-training methods (M-SRC^{Rank1}, M-SRC^{Rank}, M-SRC^{EER}). The weight-training method based on the Rank-1 IR performance presents the highest value of GAR at 0.1% FAR, reaching the rate of 76.72%. The weight-training method based on the Rank identification performance presents the lowest rate of 67.66%, and the weight-training method based on the EER performance achieves the rate of 73.97%. A one-way ANOVA (using values from 20 random splits) for GAR at 0.1% FAR across the three weight-training methods revealed that the differences in performance are statistically significant, both for the Max-Min (MM) normalization scheme $F(2, 57) = 37.36$, $p < 0.001$, and for the Z-Score (ZS) normalization scheme $F(2, 57) = 28.38$, $p < 0.001$.

Table 5. The GAR at 0.1% FAR performances in the case of the multi-source (M-SRC) fusion.

GAR at 0.1% FAR (STD) %			
Normalization	*Multi-source (multi-stimulus and multi-algorithmic) Fusion*		
	M-SRC^{Rank1}	**M-SRC^{Rank}**	**M-SRC^{EER}**
MM	**76.72** (2.67)	66.59 (4.72)	73.97 (3.83)
ZS	**76.62** (2.70)	67.66 (4.84)	73.03 (3.52)

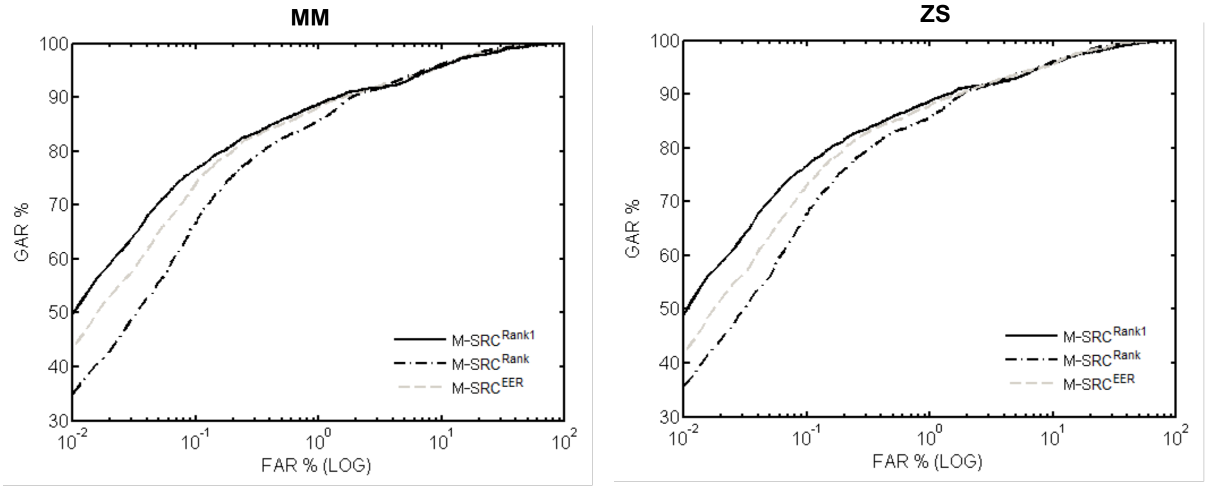

Figure 7. Comparative ROC curves using a logarithmic FAR-axis for the three weight-training methods used for the multi-source (M-SRC) fusion.

In Fig. 7, we additionally show the constructed ROC curves for the three weight-training fusion methods using a log FAR-axis. These diagrams allow for a clear investigation of the differences in the behavior of the three tested weight-training fusion methods in the important area of the low FAR values.

6. Discussion

6.1. The effects of multi-source fusion on biometric accuracy

The main objective of our research was to investigate the general effects of multi-source fusion on the eye movement-driven biometrics. For this purpose, we proposed a two stage weighted mean fusion approach, which can be used for the combination of the comparison scores generated from different algorithms under the influence of diverse visual stimuli. The suggested methodology can lead to an improved biometric performance compared to the performance achieved by each algorithm for each stimulus separately. The best achieved results of the proposed fusion methodology were a top Rank-1 IR of 88.6% (for M-SRC^{Rank1} and MM normalization), a minimal EER of 5.8% (for M-SRC^{Rank} and MM normalization), and a top GAR at 0.1% FAR of 76.7% (for M-SRC^{Rank1} and MM normalization). These results comprise a substantial improvement for the field of the eye movement-driven biometrics, and underscore the significance of the multi-source information fusion in the specific field of research.

The second objective of our research was to analyze the relative contribution of the multi-stimulus and the multi-algorithmic fusion in the overall performance. A close inspection of the results in Tables 1 to 4, reveals a clear edge of the process of multi-stimulus fusion in terms of performance improvement. Specifically, the CEM-B algorithm improves considerably its baseline rates of 47.4% Rank-1 IR and 14.4% EER, to 82% Rank-1 IR and 7.3% EER. This performance equals to a relative improvement of 72.9% for the Rank-1 IR, and a relative improvement of 51.5% for the EER. Multi-stimulus fusion also improves the performances of the other two employed biometric algorithms. For the OPC algorithm, there is a relative improvement of 32.1% for the Rank-1 IR, and of 4.9% for the EER, whereas for the FDM algorithm there is relative improvement of 13.2% for the Rank-1 IR, and 14.7% for the EER. The application of the multi-algorithmic fusion after the multi-stimulus fusion can lead to the improvement of the biometric accuracy even further. Although it is not as drastic as in the case of the multi-stimulus fusion, the additional relative improvement of 8% for the Rank-1 IR and 19.9% for the EER, is considerable.

6.2. Characteristics of the weighted mean fusion scheme

The third objective of our research was to assess the benefits of the proposed weight mean fusion scheme in comparison to other alternatives. In our case, the two-stage fusion mechanism allows training of the weights by taking into consideration the behavior of different algorithms for different stimuli. It should be also noted that the suggested scheme allows the incorporation of more than one matcher per algorithm, since it permits the utilization of different normalization functions during the two stages of fusion. In overall, the proposed weighted mean fusion scheme outperforms the tested alternatives of the Simple Mean and the Random Forests fusion, both in the identification and in the verification scenario (Tables 3 and 4). The importance of these results can be further emphasized considering the high computational cost of the Random Forests algorithm, which is a method based on ensemble learning.

The fourth objective of our research was to evaluate the efficacy of the proposed weight-training method based on Rank-1/Rank identification performance, compared to the more traditional approach based on the verification performance (EER). Such an investigation can be additionally supported by studies showing that the good verification performance does not always imply a good identification performance, and vice versa [54]. The bar diagrams showing the calculated weights (Fig. 4 and 5) demonstrate that all the evaluated weight-training methods can extract the fusion weights with satisfactory stability. This stable behavior can be attributed mainly to the large volume of training subjects used in this work. The evaluation experiments showed that the suggested method based on training with the Rank-1 IR provides the optimum biometric performance for the identification scenario, and performs similarly with the other two weight-training methods for the verification scenario. Compared to the classic EER-based weight-training method, the proposed rank identification-based method can be a more favorable solution for systems needed to operate on both modes of biometric recognition (identification and verification). Furthermore, the training procedure based on the Rank-1 IR can be also the preferable choice considering the computational cost, since it does not demand the construction of the ROC curves which are needed for the calculation of the EER.

6.3. Practical considerations and dynamic biometric scenarios

Given that the aim of the current work was to assess in a concrete way the improvements that can be achieved by the multi-source stimulus fusion compared to the baseline eye movement-driven approaches, several steps were adopted to ensure the high quality of the recorded data. In order to capture the eye movements we employed a commercial high-grade eye tracking device. Also, during data capturing we stabilized the heads of the subjects using a chinrest with a forehead. Thus, the application of the developed scheme in a more practical scenario imposes the recording of the eye movements with a relative tolerance in head movements and lighting conditions. The evolution of the eye-tracking technology already shows considerable progress to this direction, with the development of more robust remote eye trackers [55], and wearable-based eye-tracking solutions [56]. There are also attempts to make the eye-tracking technology more affordable, with the development of low-cost devices of satisfactory accuracy [57].

Another practical consideration involves the time needed to record the eye movements. Since the eye movements are evolving dynamically in time, they cannot be captured within a single frame as it can be done for the iris and the fingerprint images. Although this is a limitation of the eye movement biometrics, at the same time, the dynamic and behavioral nature of the eye movement provides unique advantages in terms of counterfeit resistance. Thus, the eye movement-driven algorithms can find application as soft biometric modules in traditional biometric systems, in order to provide anti-spoofing resistance [58, 59] and continuous identity monitoring [60].

The proposed multi-source fusion scheme can also provide practical advantages for creating more dynamic biometric recognition systems. In our method the information is combined with the use of *stimulus-specific* and *algorithm-specific* weights. Thus, the relative duration and/or the order of stimuli presentation can be dynamically chosen, allowing for the development of adaptive biometric systems. Next, we provide an example of an adaptive biometric recognition scenario. Let us assume that the user initially enrolls into the system by observing the four types of stimulus with equal time durations. In this case, if we denote with d_{tot} the total duration of presentation, the user will register $0.25 \cdot d_{tot}$ for the HOR stimulus, $0.25 \cdot d_{tot}$ for the RAN stimulus, $0.25 \cdot d_{tot}$ for the TEX stimulus, and $0.25 \cdot d_{tot}$ for the VID stimulus. Now, let us assume that during a subsequent recognition attempt, the system dynamically changes the relative duration (and/or the order) of stimuli presentation, for example $0.4 \cdot d_{tot}$ for the TEX stimulus, $0.1 \cdot d_{tot}$ for the RAN stimulus, $0.3 \cdot d_{tot}$ for the HOR stimulus, and $0.2 \cdot d_{tot}$ for the VID stimulus. In this case, the biometric system which generates the stimuli presentation can also modulate the *stimulus-specific* and *algorithm-specific* weights in response to the current presentation settings, with the aim to maximize the probability of an accurate recognition result. Inversely, let us assume that someone tries to spoof-attack the biometric system, e.g. by recording the eye movements during the initial enrollment and replay the recording during a next trial. In this case, the difference in the stimulus presentation settings during the subsequent recognition attempt will lower the possibilities of a successful spoofing attack by replaying the previously recorded eye movements, given the different modulation of the employed fusion weights.

6.4. Limitations

Our current research is subject to certain limitations. The experiments for collecting the eye movement recordings were conducted within the same day. Previous research has shown that the biometric accuracy can be affected by the appearance of template aging effects [61, 62]. Thus, an investigation with a database recorded over a longer time period would allow for an evaluation of the relative effects of the multi-source fusion for larger time intervals. Furthermore, the constructed database consists of two recordings per subject for every stimulus. Although the large number of subjects supports the stability of the extracted results, it would be interesting to investigate the behavior of the evaluated weight-training methods in the case of multiple recordings per subject. Finally, it should be noted that the proposed multi-source fusion scheme practically requires separate time for making the recording for each stimulus. However, the disadvantage of this prolonged recording duration can be partially counterbalanced by the advantages provided by the use of multiple visual stimuli in terms of the achieved performance improvement and the creation of dynamic biometric recognition scenarios.

7. Conclusion

This work investigated the effects of multi-source information fusion in the emerging field of eye movement-driven biometrics. The behavioral characteristics of the eye movements induce a certain degree of inaccuracy on the extracted features. To this context, the combination of information coming from different sources provides a useful mechanism for facilitating performance advancements in terms of recognition accuracy and robustness. In this paper, we introduced and evaluated the new concept of multi-stimulus fusion, i.e. the combination of information extracted from the eye movements while observing different visual stimuli. Additionally, we investigated the potential of the multi-algorithmic fusion by taking into consideration the existing interrelationships between the eye movement-driven algorithms and the different types of visual stimuli. Our experimental results suggest that the application of multi-source weighted fusion can lead to significant improvements in performance, when compared to the single-algorithm and single-stimulus baselines. In our future work, we plan to investigate the characteristics of the proposed fusion scheme for recordings of limited duration, and for datasets that contain multiple enrollments per subject. Also, our future research will focus on the effects of template aging on the developed fusion scheme.

Acknowledgements

This work is supported in part by the NSF CAREER Grant #CNS-1250718 and the NIST Grant #60NANB14D274. Special gratitude is expressed to T. Miller, Ch. Heinich, and N. Myers for proctoring eye movement recordings.

References

[1] A.K. Jain, P.J. Flynn, A. Ross, Handbook of Biometrics, Springer-Verlag New York, Inc., 2007.

[2] R. Brunelli, D. Falavigna, Person Identification Using Multiple Cues, IEEE Transanctions on Pattern Analysis and Machine Intelligence (PAMI), 17 (1995) 955-966.

[3] A. Ross, R. Govindarajan, Feature Level Fusion Using Hand and Face Biometrics, in: SPIE, 2005, pp. 196-204.

[4] R.N. Rodrigues, L.L. Ling, V. Govindaraju, Robustness of multimodal biometric fusion methods against spoof attacks, J. Vis. Lang. Comput., 20 (2009) 169-179.

[5] L. Hong, A. Jain, Integrating faces and fingerprints for personal identification, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 20 (1998) 1295-1307.

[6] A. Ross, A. Jain, Information fusion in biometrics, Pattern Recognition Letters, 24 (2003) 2115-2125.

[7] F. Yang, B. Ma, Q.x. Wang, D. Yao, Information Fusion of Biometrics Based-on Fingerprint, Hand-geometry and Palm-print, in: 2007 IEEE Workshop on Automatic Identification Advanced Technologies, 2007, pp. 247-252.

[8] Y. Wang, T. Tan, A.K. Jain, Combining face and iris biometrics for identity verification, in: 4th International Conference on Audio- and Video-based Biometric Person Authentication, Springer-Verlag, 2003, pp. 805-813.

[9] K. Chang, K.W. Bowyer, S. Sarkar, B. Victor, Comparison and combination of ear and face images in appearance-based biometrics, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 25 (2003) 1160-1165.

[10] S. Ben-Yacoub, Y. Abdeljaoued, E. Mayoraz, Fusion of face and speech data for person identity verification, IEEE Transactions on Neural Networks, 10 (1999) 1065-1074.

[11] R. Giot, B. Hemery, C. Rosenberger, Low Cost and Usable Multimodal Biometric System Based on Keystroke Dynamics and 2D Face Recognition, in: 20th International Conference onPattern Recognition (ICPR), 2010, pp. 1128-1131.

[12] A. Kale, A.K. Roychowdhury, R. Chellappa, Fusion of gait and face for human identification, in: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04), 2004, pp. V 901-904.

[13] A.K. Jain, S. Prabhakar, S. Chen, Combining multiple matchers for a high security fingerprint verification system, Pattern Recognition Letters, 20 (1999) 1371-1379.

[14] A. Ross, A. Jain, J. Reisman, A hybrid fingerprint matcher, in: 16th International Conference on Pattern Recognition (ICPR), 2002, pp. 795-798.

[15] R. Cappelli, D. Maio, D. Maltoni, Combining Fingerprint Classifiers, in: Multiple Classifier Systems, Springer Berlin Heidelberg, 2000, pp. 351-361.

[16] M. Vatsa, R. Singh, A. Noore, Improving Iris Recognition Performance Using Segmentation, Quality Enhancement, Match Score Fusion, and Indexing, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 38 (2008) 1021-1035.

[17] L. Xiaoguang, W. Yunhong, A.K. Jain, Combining classifiers for face recognition, in: 2003 International Conference on Multimedia and Expo (ICME '03) 2003, pp. III 13-16.

[18] A.L. Rukhin, I. Malioutov, Fusion of biometric algorithms in the recognition problem, Pattern Recognition Letters, 26 (2005) 679-684.

[19] J. Han, B. Bhanu, Gait Recognition by Combining Classifiers Based on Environmental Contexts, in: T. Kanade, A. Jain, N.K. Ratha (Eds.) Audio- and Video-Based Biometric Person Authentication, Springer Berlin Heidelberg, 2005, pp. 416-425.

[20] FBI, IAFIS, in, https://www.fbi.gov/about-us/cjis/fingerprints_biometrics/iafis/iafis, 2015.

[21] S. Prabhakar, A.K. Jain, Decision-level fusion in fingerprint verification, Pattern Recognition, 35 (2002) 861-874.

[22] J. Jang, K.R. Park, J. Son, Y. Lee, Multi-unit Iris Recognition System by Image Check Algorithm, in: D. Zhang, A.K. Jain (Eds.) Biometric Authentication, Springer Berlin Heidelberg, 2004, pp. 450-457.

[23] G.L. Marcialis, F. Roli, Fingerprint verification by fusion of optical and capacitive sensors, Pattern Recognition Letters, 25 (2004) 1315-1322.

[24] C.C. Chibelushi, J.S.D. Mason, F. Deravi, Feature-level data fusion for bimodal person recognition, in: Sixth International Conference on Image Processing and Its Applications, 1997, pp. 399-403.

[25] J. Kittler, J. Matas, K. Jonsson, M.U. Ramos Sánchez, Combining evidence in personal identity verification systems, Pattern Recognition Letters, 18 (1997) 845-852.

[26] K. Nandakumar, C. Yi, S.C. Dass, A.K. Jain, Likelihood Ratio-Based Biometric Score Fusion, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 30 (2008) 342-347.

[27] T.K. Ho, J.J. Hull, S.N. Srihari, Decision combination in multiple classifier systems, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 16 (1994) 66-75.

[28] V. Chatzis, A.G. Bors, I. Pitas, Multimodal decision-level fusion for person authentication, IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 29 (1999) 674-680.

[29] C.D. Holland, O.V. Komogortsev, Complex Eye Movement Pattern Biometrics: The Effects of Environment and Stimulus, IEEE Transactions on Information Forensics and Security, 8 (2013) 2115-2126.

[30] C.C. Collins, The human oculomotor control system, Basic Mechanisms of Ocular Motility and Their Applications, (1975) 145-180.

[31] R. Groner, M.T. Groner, Attention and eye movement control: An overview, European archives of psychiatry and neurological sciences, 239 (1989) 9-16.

[32] O.V. Komogortsev, A. Karpov, L.R. Price, C. Aragon, Biometric authentication via oculomotor plant characteristics, in: 5th IAPR International Conference on Biometrics (ICB), 2012, pp. 413-420.

[33] C.D. Holland, O.V. Komogortsev, Complex eye movement pattern biometrics: Analyzing fixations and saccades, in: 2013 International Conference on Biometrics (ICB), 2013, pp. 1-8.

[34] I. Rigas, O.V. Komogortsev, Biometric Recognition via Probabilistic Spatial Projection of Eye Movement Trajectories in Dynamic Visual Environments, IEEE Transactions on Information Forensics and Security, 9 (2014) 1743-1754.

[35] G. Westheimer, Eye movement responses to a horizontally moving visual stimulus, AMA Arch Ophthalmol., 52 (1954) 932-941.

[36] K. Rayner, Eye movements in reading and information processing: 20 years of research, Psychological Bulletin, 124 (1998) 372-422.

[37] M. Dorr, T. Martinetz, K.R. Gegenfurtner, E. Barth, Variability of eye movements when viewing dynamic natural scenes, J Vis., 10 (2010) article 28.

[38] P. Kasprowski, J. Ober, Eye Movements in Biometrics, in: D. Maltoni, A.K. Jain (Eds.) Biometric Authentication, Springer Berlin Heidelberg, 2004, pp. 248-258.

[39] R. Bednarik, T. Kinnunen, A. Mihaila, P. Fränti, Eye-Movements as a Biometric, in: H. Kalviainen, J. Parkkinen, A. Kaarna (Eds.) Image Analysis, Springer Berlin Heidelberg, 2005, pp. 780-789.

[40] T. Kinnunen, F. Sedlak, R. Bednarik, Towards task-independent person authentication using eye movement signals, in: 2010 Symposium on Eye-Tracking Research & Applications (ETRA), ACM, 2010, pp. 187-190.

[41] O.V. Komogortsev, A. Karpov, C.D. Holland, CUE: counterfeit-resistant usable eye movement-based authentication via oculomotor plant characteristics and complex eye movement patterns, in: SPIE, 2012, pp. 83711X-83711X-83719.

[42] I. Rigas, G. Economou, S. Fotopoulos, Biometric identification based on the eye movements and graph matching techniques, Pattern Recognition Letters, 33 (2012) 786-792.

[43] V. Cantoni, C. Galdi, M. Nappi, M. Porta, D. Riccio, GANT: Gaze analysis technique for human identification, Pattern Recognition, 48 (2015) 1023-1034.

[44] H.-J. Yoon, T.R. Carmichael, G. Tourassi, Gaze as a biometric, in: SPIE, 2014, pp. 903707-903707-903707.

[45] O. Komogortsev, C. Holland, A. Karpov, L.R. Price, Biometrics via Oculomotor Plant Characteristics: Impact of Parameters in Oculomotor Plant Model, ACM Trans. Appl. Percept., 11 (2014) 1-17.

[46] H. Hotelling, The Generalization of Student's Ratio, The Annals of Mathematical Statistics, 2 (1931) 360-378.

[47] T.W. Anderson, On the Distribution of the Two-Sample Cramér-von Mises Criterion, The Annals of Mathematical Statistics, 33 (1962) 1148-1159.

[48] T. Judd, F. Durand, A. Torralba, A benchmark of computational models of saliency to predict human fixations, in: MIT-CSAIL-TR-2012-001, 2012.

[49] A. Jain, K. Nandakumar, A. Ross, Score normalization in multimodal biometric systems, Pattern Recognition, 38 (2005) 2270-2285.

[50] R. Snelick, U. Uludag, A. Mink, M. Indovina, A. Jain, Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 27 (2005) 450-455.

[51] SR-Research, EyeLink 1000, in, http://www.sr-research.com/EL_1000.html, 2015.

[52] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters, 27 (2006) 861-874.

[53] Y. Ma, B. Cukic, H. Singh, A Classification Approach to Multi-biometric Score Fusion, in: T. Kanade, A. Jain, N.K. Ratha (Eds.) Audio- and Video-Based Biometric Person Authentication, Springer Berlin Heidelberg, 2005, pp. 484-493.

[54] B. DeCann, A. Ross, Can a poor verification system be a good identification system? A preliminary study, in: 2012 IEEE International Workshop on Information Forensics and Security (WIFS), 2012, pp. 31-36.

[55] SMI, RED250-RED500, in, http://www.smivision.com/en/gaze-and-eye-tracking-systems/products/red-red250-red-500.html, 2015.

[56] Tobii, Glasses 2, in, http://www.tobii.com/en/eye-tracking-research/global/landingpages/tobii-glasses-2/, 2015.

[57] THEEYETRIBE, Eye Tribe Tracker, in, https://theeyetribe.com/, 2015.

[58] O.V. Komogortsev, A. Karpov, C.D. Holland, Attack of Mechanical Replicas: Liveness Detection With Eye Movements, IEEE Transactions on Information Forensics and Security, 10 (2015) 716-725.

[59] I. Rigas, O.V. Komogortsev, Eye movement-driven defense against iris print-attacks, Pattern Recognition Letters, In Press.

[60] K. Niinuma, P. Unsang, A.K. Jain, Soft Biometric Traits for Continuous User Authentication, IEEE Transactions on Information Forensics and Security, 5 (2010) 771-780.

[61] P. Kasprowski, O.V. Komogortsev, A. Karpov, First eye movement verification and identification competition at BTAS 2012, in: IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS), 2012, pp. 195-202.

[62] O.V. Komogortsev, C.D. Holland, A. Karpov, Template aging in eye movement-driven biometrics, in: SPIE, Biometric and Surveillance Technology for Human and Activity Identification XI, 2014, pp. 90750A-90750A-90759.