IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 July 2024, accepted 24 July 2024, date of publication 9 August 2024, date of current version 20 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3437291

RESEARCH ARTICLE

DeepFace-Attention: Multimodal Face Biometrics for Attention Estimation With Application to e-Learning

ROBERTO DAZA, LUIS F. GOMEZ, JULIAN FIERREZ, (Member, IEEE), AYTHAMI MORALES, RUBEN TOLOSANA, AND JAVIER ORTEGA-GARCIA, (Fellow, IEEE)
Biometrics and Data Pattern Analytics Laboratory, Universidad Autonoma de Madrid, Campus de Cantoblanco, 28049 Madrid, Spain

Corresponding author: Roberto Daza (roberto.daza@uam.es)

This work was supported in part by project HumanCAIC under Grant TED2021-131787B-I00 MICINN; in part by project BBforTAI under Grant PID2021-127641OB-I00 MICINN/FEDER; in part by project BIO-PROCTORING (GNOSS Program, Agreement Ministerio de Defensa-UAM-FUAM dated 29-03-2022); in part by the Catedra ENIA UAM-VERIDAS en IA Responsable (NextGenerationEU PRTR) under Grant TSI-100927-2023-2; and in part by the Autonomous Community of Madrid. The work of Roberto Daza was supported by the FPI Fellowship from MINECO/FEDER. The work of Aythami Morales was supported by the Madrid Government (Comunidad de Madrid-Spain) under the Multiannual Agreement with Universidad Autónoma de Madrid in the line of Excellence for the University Teaching Staff in the context of the Regional Program of Research and Technological Innovation (V PRICIT).

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of the Autonomous University of Madrid under Application No. 11/04/2023, and performed in line with the Declaration of Helsinki.

ABSTRACT This work introduces an innovative method for estimating attention levels (cognitive load) using an ensemble of facial analysis techniques applied to webcam videos. Our method is particularly useful, among others, in e-learning applications, so we trained, evaluated, and compared our approach on the mEBAL2 database, a public multi-modal database acquired in an e-learning environment. mEBAL2 comprises data from 60 users who performed 8 different tasks. These tasks varied in difficulty, leading to changes in their cognitive loads. Our approach adapts state-of-the-art facial analysis technologies to quantify the users' cognitive load in the form of high or low attention. Several behavioral signals and physiological processes related to the cognitive load are used, such as eyeblink, heart rate, facial action units, and head pose, among others. Furthermore, we conduct a study to understand which individual features obtain better results, the most efficient combinations, explore local and global features, and how temporary time intervals affect attention level estimation, among other aspects. We find that global facial features are more appropriate for multimodal systems using score-level fusion, particularly as the temporal window increases. On the other hand, local features are more suitable for fusion through neural network training with score-level fusion approaches. Our method outperforms existing state-of-the-art accuracies using the public mEBAL2 benchmark.

INDEX TERMS Attention estimation, behavioral analysis, cognitive load, deep learning, e-learning, eyeblink, facial action units, head pose detection, heart rate detection, multi-modal learning.

I. INTRODUCTION

Attention is defined as the ability to focus, specifically, to exert on a conscious cognitive effort regarding a specific task or stimulus at a given moment [1], [2]. Therefore, it is used as a measure of the exerted effort. The level of attention can vary from a state of high attention, where a person is highly concentrated and experiences high levels of cognitive load and mental effort, to low levels, where a person is distracted or uninterested.

Attention estimation has proven to be of great value in important areas such as driver fatigue detection [3], [4],

The associate editor coordinating the review of this manuscript and approving it for publication was Vincenzo Conti.

FIGURE 1. Examples of different real students' attention levels during an e-learning session. (Top) High attention image sequence. (Bottom) Low attention image sequence.

advertising and product design [5], mental health disorders [6], lie detection [7], [8], human-computer interfaces [9], education [10], etc.

Attention estimation is particularly valuable in e-learning environments [11], [12] because it offers feedback on students' cognitive and emotional states during online sessions. This is significant as attention is defined as the cognitive effort exerted on a task [1] and plays a pivotal role in ensuring accurate comprehension during learning. In e-learning environments, there are challenges compared to face-to-face education, with one of the most important being the lack of direct contact between the teacher and the student. This results in the teacher being unaware of the student's study difficulties, like high or low levels of attention. Video-based attention estimation technologies overcome this limitation [13], representing a valuable tool to enhance both face-to-face and online education.

Facial gestures often provide subtle indicators of an individual's attention level or cognitive load. When people are intensely focused or experiencing high cognitive demands, their facial expressions can change, reflecting the strain or concentration they are undergoing (see Fig. 1). Automatic attention estimation through image processing is a challenging task still under development. In this regard, the recent advances in face analysis techniques based on deep learning have also helped to improve attention estimation based on computer vision methods. The most advanced multimodal systems for attention estimation have reached around 80% accuracy, outperforming the majority of existing monomodal systems [12], [14]. Multimodal systems stand out for considering multiple variables that affect attention in the learning process, which allows a more global and complete perspective [15].

Taking into consideration all of the above, the main contributions of the present paper are:

- We present a novel multimodal learning framework for attention estimation through image processing. This framework performs facial analysis to relate high and low levels of attention with behavior and physiological processes such as eyeblink, face gestures, and head pose, among others.
- Our framework consists of 5 modules built on Convolutional Neural Networks (CNNs) that are trained to extract facial features that potentially correlate with attention. The most relevant modules for attention estimation and their effective combinations are identified within the e-learning context of the mEBAL2 database.
- The results indicate that in multimodal attention estimation systems using score fusion, global features provide additional discriminating information compared to local features. However, multimodal attention estimation systems based on score fusion with neural network training generalize better with local features.
- Our approach outperforms the state of the art, achieving a classification accuracy in attention level estimation of 85.92% on the mEBAL2 database.

A preliminary version of this article was presented in [12]. This article significantly improves [12] in various aspects:

- Compared to MATT [12], we now add a new module for heart rate estimation and study its relationship with attention estimation.
- The mEBAL2 [16] database for attention estimation is used to train and evaluate the proposed system. In comparison with MATT [12] (which used the first mEBAL version with 22 users [17]), we now use the new version, mEBAL2, including 60 students with approximately 1800 minutes of video recordings. This represents a significant increase, with around 1140 additional minutes of recordings in comparison to [12].
- We add new comprehensive experiments including analysis of global and local features for each facial module. We introduce a new method of score-level fusion through neural network training and a new architecture based on feature selection.
- Unlike MATT [12], which utilized a one-minute time frame, we explored three time windows of 30, 60, and 120 seconds.
- Finally, our method outperforms the method presented in MATT [12], achieving an error reduction of 28.5% in the mEBAL2 database.

The rest of the paper is organized as follows. Section II summarizes works related to attention level estimation. Section III describes the materials and methods, including the database, proposed technologies and features to estimate attention levels. Section IV presents the experiments and comparison with other state-of-the-art approaches. Finally, section V provides conclusions and future investigations.

II. RELATED WORK

A. BRAIN ACTIVITY MEASUREMENT

Attention estimation has been widely studied and currently there are different methods that come along with certain benefits and limitations [18]. Some of the most popular ones are:

1) ELECTROENCEPHALOGRAPHY (EEG)

The EEG records the electrical activity of the brain through electrodes placed on the scalp. It measures neural activity by detecting changes in the voltage fluctuations generated by brain cells, specifically, the ones produced usually by synaptic excitations of the dendrites of pyramidal cells in the top layer of the brain cortex [19], [20]. The strength of the signals primarily relies on the synchronized firing of numerous neurons and fibers. Thousands or even millions of neurons are required to capture information effectively [18]. EEG data is recognized as one of the most efficient and unbiased approaches in estimating attention levels [21], [22], since these signals are sensitive to mental effort, cognitive demands, and mental states such as learning, deception, perception, and stress. Therefore, EEG provides real-time information about brain activity and it's particularly useful for capturing quick changes in attention. EEG can be condensed to of five different signal types that reflect different mental states and activities. These signals are classified into different frequency bands: δ ($<$ 4Hz), θ (4-8 Hz), α (8-13 Hz), β (13-30 Hz), and γ ($>$ 30 Hz). However, the main disadvantage of this method is its intrusiveness, requiring precise tools to be placed on the student's head, which becomes impractical in e-learning environments with thousands of students.

2) PHYSIOLOGICAL

This category associates attention with physiological responses like heart rate [23], [24], eyeblink [11], [12], [16], [17], eye pupil size [25], [26], electrodermal activity [27], etc. To measure these physiological signals and then correlate them with attention, specific sensors are used for each method that are then combined to obtain higher accuracy in attention estimation.

3) BEHAVIOR

In comparison with the physiological category, this category analyzes the user's noticeable patterns and behaviors to deduce attention levels. It's based on external behavior observation that has proven to have a close relation with attention. Some of these behaviors are head pose [14], [28], [29], [30], gaze tracking [31], [32], facial expressions [33], [34], [35], physical actions that happen to be related with attention (e.g., leaning closer to the screen) [36], etc.

B. ATTENTION ESTIMATION METHODS BASED ON IMAGE PROCESSING

Here we employ images obtained from the webcam to infer the attention level of the users. The main advantage of this approach is that it doesn't require specialized sensors more than a webcam, which makes it particularly attractive in areas like education, where accessibility is important. Currently there are monomodal systems like ALEBK [11] and multimodal ones like MATT [12]. For example, MATT combines physiological and behavior estimations (pulse, facial analysis, etc). Multimodal systems have proven to be more efficient in attention estimation.

The article [37] proposed 2 monomodal methods to detect cognitive load in car driving environments. The used database defined 3 states of cognitive load (high, medium, and low), which corresponded to variable difficulty activities (based on n-back task) that drivers had to perform; and the database had a total of 92 users. The proposed methods were based on the eye state, starting with the first method that focused on the eye pupil's position estimation (using face detection, landmark detection, etc) with Hidden Markov Models (HMMs) to estimate cognitive load. The second approach was based on Convolutional Neural Networks with 7 convolutional layers, and the input was a temporally-stacked sequence of raw grayscale eye region images. The HMMs approach reached an average precision of 77.7% while the CNN got 86.1%. The main issue was the cognitive load assumption without validating it using specific sensors, like EEG for example.

ALEBk [11] represents a monomodal approach based on the relation between eyeblink and cognitive activity. Several studies have found clear evidence [11], [17], [38], [39] that lower eyeblink rates are associated with high attention levels, and vice versa. Based on this assumption, ALEBk [11] uses an eyeblink detector supported by convolutional neural networks to obtain the eyeblink frequency using RGB videos. With this information, the system classifies between high or low attention. The network was trained using the mEBAL [17] database with 22 users performing tasks in an e-learning environment. Attention ground truth was obtained with an EEG band and the system reached a maximum accuracy (1-EER) of 70% approximately.

The multimodal approach presented in [14] used a Kinect One sensor to perform attention estimation. It only used behavior features, specifically gaze point, body posture and facial movements. The features were obtained from the signals of the Kinect SDK. This process included normalizing and filtering the signals using z-scores and an 11s-wide Gaussian filter. Subsequently, a 7-feature vector was selected by combining these signals. Finally, a 3-level attention classification was made (low, medium, high) using different classifiers like decision tree, K-nearest neighbors, Subspace K-NN, etc. This study used a database captured in an e-learning environment of 18 users with a length of 122 minutes in total. The way how the attention level ground truth was obtained is the main problem of this database, since it was through human observers, which can generate a lack of reliability in the results. Obtained results show a maximum accuracy of 75% with a considerable variability between users.

In [13], a multimodal system is presented to estimate attention in a learning environment. This system extracted features from the face and also head movements, like mouth features (speaking or smiling), eye aspect ratio [40], leaning closer to the screen, etc, to estimate attention. It's a simple system that uses a landmark detector to obtain the previously mentioned features from facial landmarks. Then, statistical

TABLE 1. mEBAL2 database: sensors.

Sensors	Sampling Rate
EEG Band[1]	1 Hz
1 RGB[2] camera	30 Hz
2 NIR[2] cameras	30 Hz

measures like max, min, mean, variance, range and spectral entropy of face and head features are used for a random forest regression model, that predicts mean attention in a 10-second window. The used database consisted of recorded videos (176 minutes) of 7 middle school students while they interacted with an online tutoring system, along with EEG data. The authors reported an average RMSE of 12.66 and indicated that both face and head movements provided useful information for attention estimation.

The authors of [41] proposed a multimodal attention estimation system for classrooms with several students to improve learning. The artificial vision approach used features like head pose, gaze direction and facial expression (facial action units) obtained with OpenFace [42] and regression models to estimate attention were trained with them. The approach classified the student's commitment level as ''attentive'' and ''non-attentive'' in one-second time frames. The database was obtained from university seminars with 52 students recorded with 3 cameras, even though the automatic approach used only 30 users. The attention level labels used as ground truth were obtained by evaluators that observed each student's behavior throughout sessions. The head pose feature got the least correlation regarding manual scores, and the highest correlation was reached with the combination of all 3 modules (r=0.61).

MATT [12] represents a multimodal approach that uses a simple webcam and it's based on different Convolutional Neural Network modules that extract behavior and physiological features (head pose, eyeblink, facial action units, etc). For each module, a Support Vector Machine (SVM) is used as a binary classifier to determine high or low attention levels and at the end, all modules are combined with a score sum. Similar to ALEBk [11], this approach was trained and evaluated on mEBAL [17] database with 22 users and obtained a maximum accuracy (1-EER) of 82% approximately.

C. MULTIMODAL MACHINE LEARNING

Multimodal systems have demonstrated great potential to improve the performance of unimodal systems [15], [43], due to their enhanced comprehension capabilities. By integrating various data sources, these systems can leverage the redundancy and complementarity of information to achieve more accurate and robust results. Specifically, in attention estimation, analyzing a single facial feature category is typically not discriminative enough to classify attention levels [11], [12], [13], [14]. In contrast, multimodal systems show superior performance in estimating attention by integrating different unimodal systems based on diverse facial categories, such as eyeblinks and heart rate [12], [13], [14]. Various fusion strategies have been proposed in the literatures [15], [16], [43], [44], [45], [46], and [47], including feature level fusion, score level fusion, and model level fusion. Feature level fusion involves combining data or signals at the feature level before they are input into a classification or regression model [15], [43], [44], [48]. Leng et al. [44] employed Dual-Source Discriminative Power Analysis (DDPA) to assess the discriminative power of features from two different information sources, based on inter-class and intra-class variation, and subsequently fused them. Score level fusion, on the other hand, involves combining outputs from multiple models to reach a final decision. Various strategies are employed, including score sum, weighted sum, and voting, among others [15], [45], [48]. Other works [16], [46], [47] have implemented model level fusion. Yao et al. [47] proposed an extension of the conventional Vision Transformer (ViT). This approach applied a strategy for fusing through a structure that integrates extended visual transformers and Cross-Modality Attention (CMA), thus incorporating modality fusion directly into the model processing stages.

III. MATERIALS AND METHODS

A. DATABASE

To carry out this study, we selected the public database mEBAL2 [16], a Multimodal Database for EyeBlink Detection and Attention Level Estimation. It's the first database that we're aware of being captured in an e-learning environment, providing information on attention levels and eyeblink samples. mEBAL2 is a public database obtained in a real e-learning environment using the research platform edBB [10], [32], [49]. We used this database, which includes data from 60 students who performed various carefully designed tasks to induce changes in cognitive load. These tasks were designed to induce changes in students' attention and evaluate the cognitive load associated with each situation. Among the tasks included in the acquisition protocol, the task of committing fraud/copying was included, as previous research demonstrated that this activity requires a higher cognitive load [8]. Students were presented with diverse scenarios to engage in copying responses, like using different electronic devices (mobile phones, laptops), employing ''cheat sheets,'' interacting with peers to obtain answers, and more. The database also induced an altered state in the students, to observe how it affected their attention during the e-learning session. During a specific moment, students engage in physical exercise, inducing an altered state that affects their heart rate, simulating a state of nervousness/stress. Afterward, they resume the session.

[1]https://store.neurosky.com/pages/mindwave

[2]https://www.intelrealsense.com/wp-content/uploads/2020/06/Intel-RealSense-D400-Series-Datasheet-June-2020.pdf

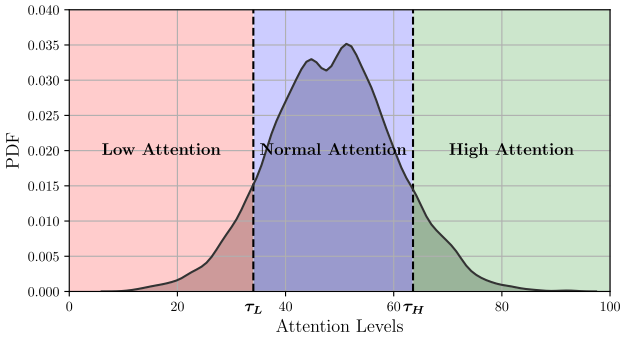


FIGURE 2. Probability density function of obtained attention with EEG band from 60 students in the mEBAL2 database [16], along with our attention levels classification (high, normal, low) with used thresholds (τ_L, τ_H).

mEBAL2 contains signals from multiple sensors, including face video and electroencephalogram (EEG) data. The data was captured with the following sensors (see Table 1): An Intel RealSense composed of 1 RGB camera and 2 NIR cameras, along with an EEG band provided by NeuroSky. It is worth mentioning that previous studies have also utilized this EEG headset to gather EEG and attention signals [20], [50], [51], as EEG measurement is considered one of the most effective methods for attention estimation. The information from the EEG band includes 5 EEG signals (δ, θ, α, β, γ). Through the official NeuroSky SDK, mEBAL2 includes information regarding attention and meditation level, and a temporal sequence with eyeblink strength. Attention and meditation levels are assigned values ranging from 0 to 100. We employed the attention levels acquired from the EEG headset as ground truth to both train and evaluate our image-based attention level estimation approach. Additionally, mEBAL2 [16] provides 10550 eyeblink samples, the largest existing public eyeblink database for research.

To summarize, mEBAL2 [16] includes data from 60 students who participated in e-learning sessions that lasted between 15 to 30 minutes. These sessions consisted of various activities related to mental load, visual attention, etc., such as filling in registration forms, answering logical and multiple-choice questions, performing visual exercises (describing images, finding differences), and more. Additionally, some of the students took part in events related to changes in attention, such as fraud/copying, physical exercise (see [49] for a video demonstration[3]). All participants gave written informed consent. The study is in accordance with the Declaration of Helsinki.

Fig. 2 shows the Probability Density Function of the attention levels of the 60 students, with an average attention level around 50%, and the most frequent attention level being 55%.

B. FACE ANALYSIS MODULES

Our proposed DeepFace-Attention estimates attention through the facial analysis of images captured by a webcam.

[3]https://www.youtube.com/watch?v=JbcL2N4YcDM

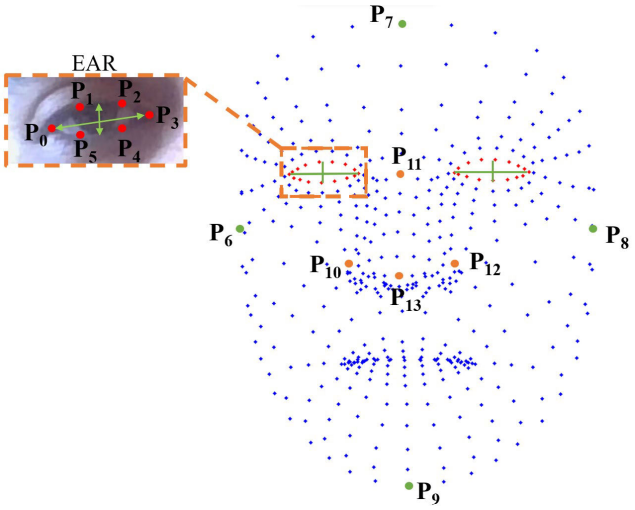


FIGURE 3. Feature extraction from the landmark detection module. On the right eye, we show Eye Aspect Ratio (EAR) calculations. We also display the landmarks used to extract the width and height of the nose and head.

Different modules based on convolutional networks are used to extract facial features based on behavior as well as physiological signals, which have proven to estimate attention [12], [14], [16], [52]. Fig. 4 shows our proposed system of attention estimation. The used modules are as follows:

1) FACE DETECTION MODULE

Our approach detects 2D facial images using a state-of-the-art RetinaFace Detector [53]. This robust single-stage face detector was trained using the Wider Face dataset [54]. Once the facial position in the image is obtained, it is used as input for the subsequent modules.

2) LANDMARK DETECTION MODULE

We use the SAN landmark detector [55] to acquire facial landmarks, which comprises a 68-landmark detection system based on VGG-16 plus 2 convolutional layers trained on the 300-W dataset [56]. The facial landmarks serve as a dual purpose in our approach. Firstly, these landmarks are used to extract facial features that have demonstrated relevance in attention estimation. Secondly, they are employed to locate the eye region of interest, which subsequently serves as input to the EyeBlink module.

Through facial landmarks, we obtain features related to attention estimation. Firstly, we focus on the eye state, specifically the Eye Aspect Ratio (EAR) [40] for each eye, which is related to the eye opening.

The EAR is calculated following the next equation:

$$\text{EAR} = \frac{\|\mathbf{P}_1 - \mathbf{P}_5\| + \|\mathbf{P}_2 - \mathbf{P}_4\|}{2\,\|\mathbf{P}_0 - \mathbf{P}_3\|} \tag{1}$$

where $\mathbf{P}_0, \ldots, \mathbf{P}_5$ are the eye landmarks shown in Fig. 3. The denominator is multiplied by 2 because only one distance is calculated for horizontal eye landmarks.

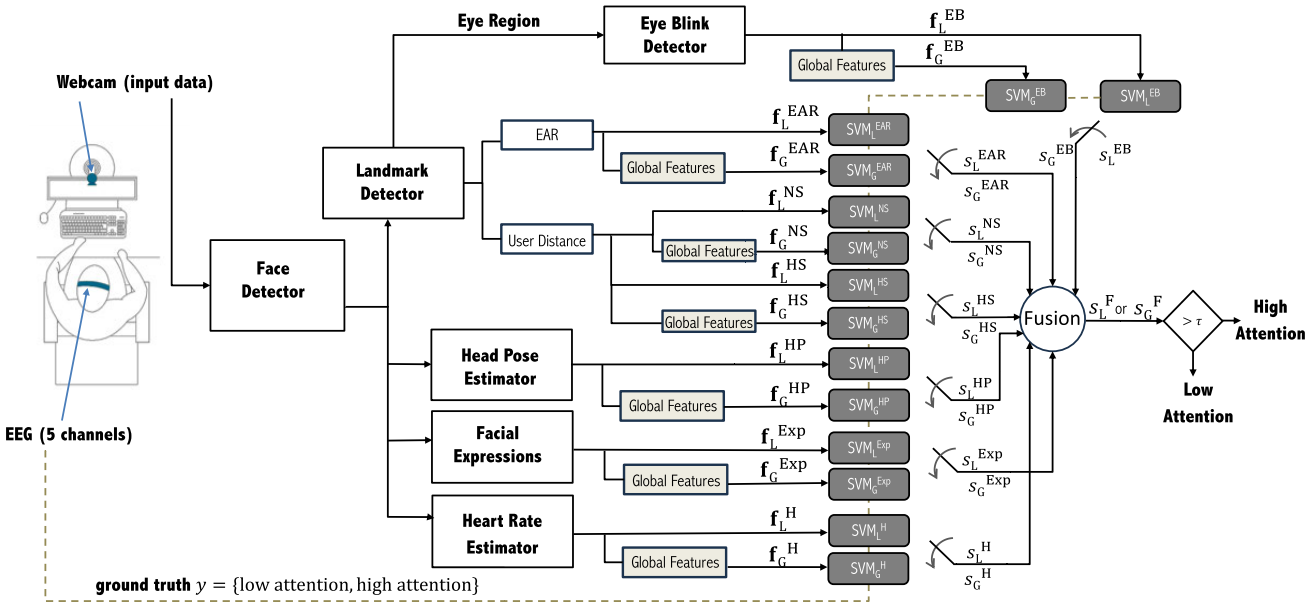


FIGURE 4. Block diagram of the proposed multimodal approach for attention estimation (DeepFace-Attention). The dashed line represents the ground truth used for training the SVMs. The two strategies used, global features ($\mathbf{f}_G$) and local features ($\mathbf{f}_L$), are shown. The feature vectors from each module are denoted as $\mathbf{f}_x^y$, and the score for each SVM is denoted as s_x^y. Here, $x \in \{L, G\}$ specifies whether the features are global or local, and y represents the facial feature category, $y \in \{EB, HP, EAR, \ldots\}$. Finally, s^F represents the fusion of scores.

We calculate the EAR parameter for each eye, so, two EAR features are obtained per frame.

The other 4 features are related to the student's distance from the screen, as previous studies have shown its usefulness in attention estimation [13]. We obtain the Width and Height of the Head and the Nose by simply subtracting the following landmarks:

$$H_W = P_{8_x} - P_{6_x} \tag{2}$$
$$H_H = P_{9_y} - P_{7_y} \tag{3}$$
$$N_W = P_{12_x} - P_{10_x} \tag{4}$$
$$N_H = P_{13_y} - P_{11_y} \tag{5}$$

where $\mathbf{P}_6, \ldots, \mathbf{P}_{13}$ are the eye landmarks shown in Fig. 3.

Finally, we normalize all the values using z-score [57], resulting in four features for each frame corresponding to the facial feature categories of Head Size (HS) and Nose Size (NS).

This landmark processing is in line with our previous works, see [58] and [59] for more details.

3) HEAD POSE ESTIMATION MODULE

The head pose is estimated using 2D facial images obtained from the facial detection module. To achieve a balance between speed and precision, we used a Convolutional Neural Network (ConvNet) based on [60]. This head pose estimator was trained with data from the Pointing 04 [61] and Annotated Facial Landmarks in the wild [62] databases. This architecture calculates the vertical (pitch) and horizontal (yaw) angles, enabling us to infer the 3D head pose from 2D facial images. This module obtains the two angles that define the 3D head pose for each frame, forming the facial feature category Head Pose (HP).

4) EYEBLINK DETECTION MODULE

The eye state has proven to be one of the most relevant indicators for attention estimation. We use an eye state classifier on each RGB frame, distinguishing between "open" or "closed" states, which is commonly employed as a blink detector in frame sequences. Our architecture is based on the approach presented in ALEBk [11], and we trained it from scratch using the mEBAL database [17], with RGB images only. The output values range between 0 and 1 and the input consists of two cropped images of the right and left eye. We apply the following approach to obtain the region of interest: *i)* face detection, *ii)* landmark detection, *iii)* face alignment using the Dlib library, *iv)* data quality assessment: we use the detectors' probabilities to evaluate the ROI quality from which we decide to maintain or not the alignment, or discarding the frame, and *v)* eye cropping: we crop the region of each eye and resize it to 50 × 50. This module obtains a value between 0 and 1 as a feature per frame for the facial feature category EyeBlink (EB).

5) FACIAL EXPRESSION MODULE

This module is based on the work by Zhang et al. [65], who created a new architecture based on the subtraction of two embeddings to extract a disentangled feature space where the facial expression embedding was compacted, and the user's identity was ignored. The two branches are two FaceNet-Inception architectures pretrained with VGGFace2,

TABLE 2. Features extracted from the face analysis modules for our proposed system. W_l is the time window size (in seconds) analyzed to extract global or local features.

Modules	Feature Categories	Local Feature Vectors	Global Feature Vectors
Landmark	EAR	$\mathbf{f}_{\mathrm{L}}^{\mathrm{EAR}} \in \mathbb{R}^{2\times W_l}$	$\mathbf{f}_{\mathrm{G}}^{\mathrm{EAR}} \in \mathbb{R}^{2\times 28}$
	HS	$\mathbf{f}_{\mathrm{L}}^{\mathrm{HS}} \in \mathbb{R}^{2\times W_l}$	$\mathbf{f}_{\mathrm{G}}^{\mathrm{HS}} \in \mathbb{R}^{2\times 28}$
	NS	$\mathbf{f}_{\mathrm{L}}^{\mathrm{NS}} \in \mathbb{R}^{2\times W_l}$	$\mathbf{f}_{\mathrm{G}}^{\mathrm{NS}} \in \mathbb{R}^{2\times 28}$
Head Pose	HP	$\mathbf{f}_{\mathrm{L}}^{\mathrm{HP}} \in \mathbb{R}^{2\times W_l}$	$\mathbf{f}_{\mathrm{G}}^{\mathrm{HP}} \in \mathbb{R}^{2\times 28}$
EyeBlink	EB	$\mathbf{f}_{\mathrm{L}}^{\mathrm{EB}} \in \mathbb{R}^{1\times W_l}$	$\mathbf{f}_{\mathrm{G}}^{\mathrm{EB}} \in \mathbb{R}^{1\times 28}$
Facial Expression	Exp	$\mathbf{f}_{\mathrm{L}}^{\mathrm{Exp}} \in \mathbb{R}^{16\times W_l}$	$\mathbf{f}_{\mathrm{G}}^{\mathrm{Exp}} \in \mathbb{R}^{16\times 28}$
Heart Rate	H	$\mathbf{f}_{\mathrm{L}}^{\mathrm{H}} \in \mathbb{R}^{1\times W_l}$	$\mathbf{f}_{\mathrm{G}}^{\mathrm{H}} \in \mathbb{R}^{1\times 28}$

TABLE 3. Description of the g_n^k ($k = 1, \ldots, 28$) global features of the global vector $\mathbf{g}_n$ extracted from each time series used in this work. Adapted from [63] and [64].

#	Feature Description	#	Feature Description
g_n^1	Total positive velocity $\sum_{(\mathbf{v}>0)}$	g_n^2	Total negative velocity $\sum_{(\mathbf{v}<0)}$
g_n^3	(1st maximum location in $\mathbf{x}$)	g_n^4	(2nd maximum location in $\mathbf{x}$)
g_n^5	(3rd maximum location in $\mathbf{x}$)	g_n^6	(average velocity $\tilde{\mathbf{v}}$)/ $\lvert\mathbf{v}\rvert_{\max}$
g_n^7	(average velocity $\tilde{\mathbf{v}}$)/$\mathbf{v}_{\max}$	g_n^8	(RMS velocity $\mathbf{v}_{\mathrm{RMS}}$)/ $\lvert\mathbf{v}\rvert_{\max}$
g_n^9	(RMS centripetal acceleration $\mathbf{a}_{c_{\mathrm{RMS}}}$)/ $\lvert\mathbf{a}\rvert_{\max}$	g_n^{10}	(RMS tangential acceleration $\mathbf{a}_{t_{\mathrm{RMS}}}$)/ $\lvert\mathbf{a}\rvert_{\max}$
g_n^{11}	(RMS acceleration $\mathbf{a}_{\mathrm{RMS}}$)/ $\lvert\mathbf{a}\rvert_{\max}$	g_n^{12}	(average abs. centripetal acceleration $\lvert\tilde{\mathbf{a}}_c\rvert$)/ $\lvert\mathbf{a}\rvert_{\max}$
g_n^{13}	standard deviation of velocity $\sigma_{\mathbf{v}}$	g_n^{14}	standard deviation of acceleration $\sigma_{\mathbf{a}}$
g_n^{15}	average abs. jerk $\lvert\tilde{\mathbf{j}}\rvert$	g_n^{16}	average jerk $\tilde{\mathbf{j}}$
g_n^{17}	maximum abs. jerk $\lvert\mathbf{j}\rvert_{\max}$	g_n^{18}	maximum jerk $\mathbf{j}_{\max}$
g_n^{19}	RMS jerk $\mathbf{j}_{\mathrm{RMS}}$	g_n^{20}	(time of $\lvert\mathbf{j}\rvert_{\max}$)
g_n^{21}	(time of $\tilde{\mathbf{j}}_{\max}$)	g_n^{22}	Total sign changes of $\mathbf{v}$
g_n^{23}	$(\sum_{(\mathbf{v}>0)} \lvert\mathbf{v}\rvert)/(\sum_{(\mathbf{v}<0)} \lvert\mathbf{v}\rvert)$	g_n^{24}	$(\sum_{(\mathbf{v}>0)})/(\sum_{(\mathbf{v}<0)})$
g_n^{25}	$\mathbf{x}_{\max} - \mathbf{x}_{\min}$	g_n^{26}	$\tilde{\mathbf{v}}/(\mathbf{x}_{\max} - \mathbf{x}_{\min})$
g_n^{27}	(Total of local maximum in $\mathbf{x}$)	g_n^{28}	(average acceleration $\lvert\tilde{\mathbf{a}}\rvert$)

where the first branch is fixed to preserve the identity information and the second branch is retrained with Google Facial Expression Comparison (FEC) dataset [66] to improve the facial expression features. The model follows the same experimental protocol proposed in [65] using the triplet loss function to obtain the disentangled facial expression space. The result is 16 features per frame for the facial feature category of Facial Expression (Exp).

6) HEART RATE DETECTION MODULE

We employ the DeepPhys model to estimate the human heart rate using remote photoplethysmography (rPPG) based on the facial video sequences. This model is based on the Convolutional Attention Network created by Chen and McDuff in [67] and implemented by Hernandez-Ortega et al. in [68], where the DeepPhys architecture was trained on the COHFACE database [69]. The model comprises two parallel Convolutional Neuronal Networks branches that extract temporal and spatial information from videos: (i) Motion branch designed to realize a short-time video analysis to detect pixel changes over the scene, and (ii) Appearance branch designed to create attention masks based on the subject's appearance to help the motion model. This module outputs $\mathbf{f}^{\mathrm{H}} \in \mathbb{R}^{1\times W_l}$, which corresponds to a heart-rate estimation every second of the time window at hand (of size W_l seconds).

C. FEATURE EXTRACTION APPROACHES: LOCAL VS GLOBAL

Considering that the analysis of long temporal sequences increases the complexity of classification algorithms based directly on the time sequences, here we study to what extent are useful and efficient global features that integrate the information across time. To integrate the temporal information from the video sequences, we have adapted the global features proposed in [63] and [70].

The face analysis modules presented in the previous section are used to extract local and global features (see Table 2). We then apply two different feature processing approaches for the extraction of local and global relationships.

First, to characterize the local relations we use the method presented in MATT [12]. The features obtained from each module, denoted as $\mathbf{f}_{x,y}$, where $x \in \{1, \ldots, N\}$ represents the specific feature and y represents the facial feature category $y \in \{\mathrm{EB}, \mathrm{HP}, \mathrm{EAR}, \ldots\}$, are used to obtain local feature vectors. For each facial feature category, a local feature vector is generated, capturing the changes in the facial attributes for high and low attention, as follows: *i)* the facial analysis

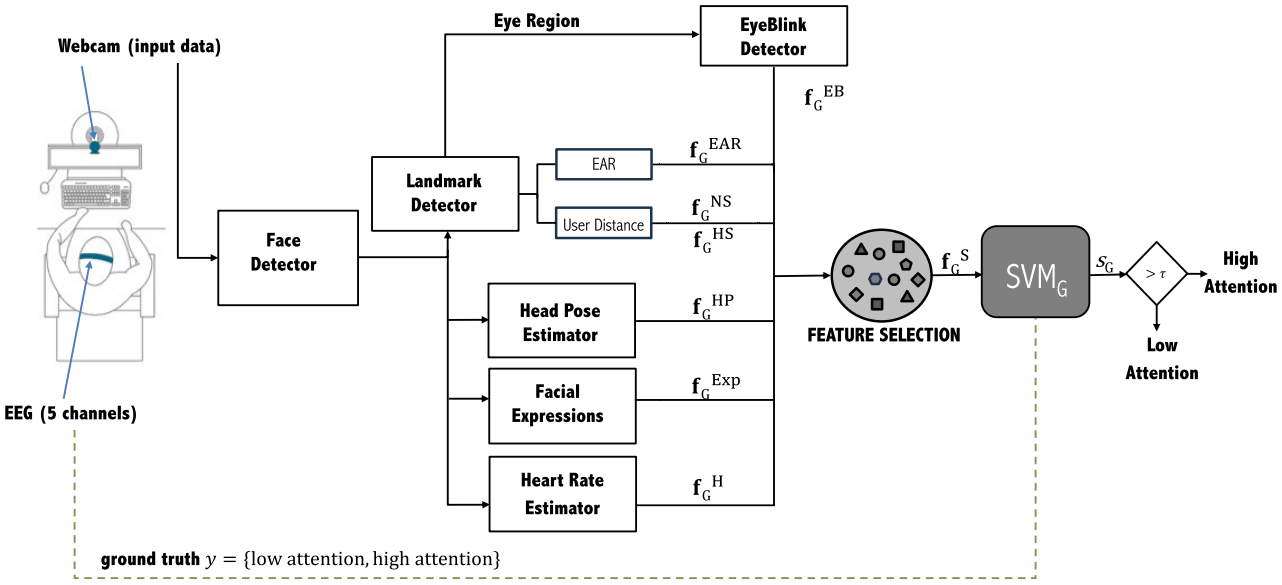


FIGURE 5. Block diagram using an approach of selection and fusion of global features for attention estimation. The dashed line represents the ground truth used for training the SVM. The global feature vector is denoted as $\mathbf{f}_{\mathrm{G}}^{y}$, where y represents the facial feature category, $y \in \{\mathrm{EB}, \mathrm{HP}, \mathrm{EAR}, \ldots\}$. $\mathbf{f}_{\mathrm{G}}^{\mathrm{S}}$ represents the vector of selected global features. Finally, the score obtained from the SVM is denoted as s_{G}.

module's features $\mathbf{f}_{x,y}$ are averaged for each second of video, generating $\bar{\mathbf{f}}_{x,y}$, and *ii)* for each facial feature category, a local feature vector $\mathbf{f}_{\mathrm{L}}^{y} \in \mathbb{R}^{N \times W_l}$ is obtained by concatenating the 1s averages $\bar{\mathbf{f}}_{x,y}$ across the time window of size W_l (30, 60, or 120 seconds), making $N \times W_l$ the dimension of the vector, where N is the number of features per second. These local feature vectors are used to estimate the attention level every second.

Second, the characterization of global relationships proposed in this work (one of the novelties here in DeepFace-Attention with respect to MATT [12]) involves extracting statistical features from the outputs of the face analysis modules, which have previously demonstrated their effectiveness in other classification tasks [63], [70]. For each facial feature category, a global feature vector $\mathbf{f}_{\mathrm{G}}^{y}$ is extracted from a sequence of features $\mathbf{f}_{\mathrm{L}}^{y} \in \mathbb{R}^{N \times W_l}$, where W_l is the time window size (in seconds) and N is the number of features per second. This sequence $\mathbf{f}_{\mathrm{L}}^{y}$ is formed as before in the local representation by concatenating the 1s averages $\bar{\mathbf{f}}_{x,y}$. The global feature vector $\mathbf{f}_{\mathrm{G}}^{y}$ for each feature category $y \in \{\mathrm{EB}, \mathrm{HP}, \mathrm{EAR}, \ldots\}$ is now defined as a set $\mathbf{g}_n \in \mathbb{R}^{28}$ with $n \in \{1, 2, \ldots, N\}$ where N is the number of features per second as described in Table 3.

D. ATTENTION LEVEL ESTIMATION BASED ON FACIAL FEATURES

Based on the facial features presented in previous sections, we propose a binary classifier to estimate periods of high or low attention.

The attention levels in the mEBAL2 dataset range from 0 to 100; however, for our study, we performed binary classification (high, low). Additionally, the attention levels vary for each student. To address these aspects, we followed the protocol proposed by ALEBk and MATT [11], [12], where two thresholds were defined for high and low attention periods segmentation: high attention (attention higher than a threshold τ_H) and low attention (attention lower than a threshold τ_L). In our case, the thresholds were obtained through the probability density function (PDF) of the attention levels from the 60 students (see Fig. 2). Specifically, we considered low attention as the values below the 10th percentile (τ_L) and high attention as the values above the 90th percentile (τ_H), as these percentiles have been shown in previous works to be separable in high and low attention.

The attention levels from the EEG band are provided every second (1 Hz). However, our approach focuses on longer temporal windows to gather enough behavioral and physiological features that can effectively classify attention. Specifically, here we study three different sliding windows of 30s, 60s, and 120s. This means that attention was estimated every second, based on the characteristics extracted from the frame sequence within the time window of size W_l seconds.

We then calculated the band attention level per window (reducing the impact of possible errors and obtaining a more accurate value of the captured attention by the band) and assigned a high or low label. After obtaining the labels, we analyzed the video sessions using all modules. For each facial feature category, we generated two vectors of both local and global features for the applied windows.

We trained two Support Vector Machine (SVM) binary classifiers for each facial feature category, one using local and other using global features as described in Section III-C (see Fig. 4).

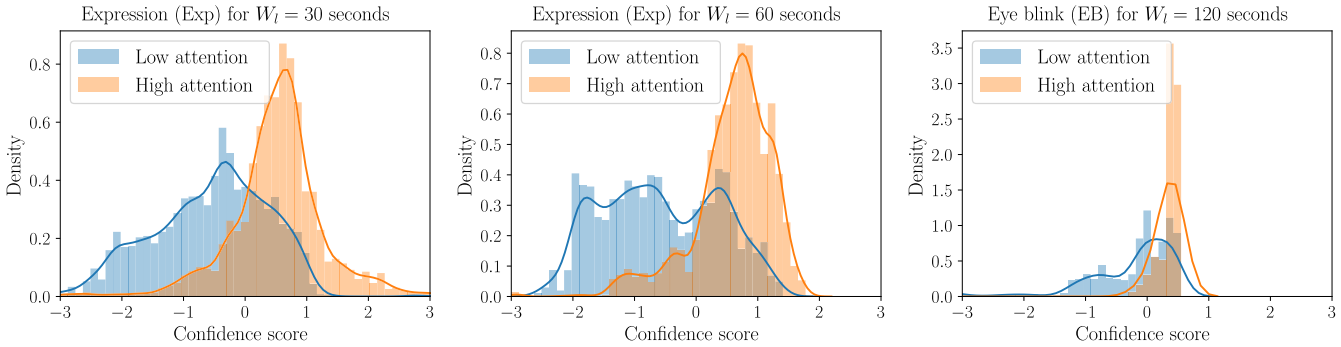


FIGURE 6. Probability density distributions of the confidence scores obtained by our attention estimation systems for the best approach in each of the three time windows considered (from left to right: 30s, 60s, and 120s) using local features. In order to simplify the performance analysis/comparison, our experimental discussion is focused on binary classification into low/high attention using the score threshold that maximizes classification accuracy.

All SVMs were trained with a linear kernel, employing a squared L2 penalty with a regularization hyper-parameter C ranging from $1e^{-8}$ to $1e^{2}$ with steps in powers of 10. Additionally, a tolerance of $1e^{-3}$ is set for the stopping criterion. It is important to mention that this work also evaluated the performance of RBF kernel SVM and Random Forest. However, the differences in the performance of the three proposed algorithms were marginal. For greater clarity, the paper only presents the results of the linear SVM classifiers.

To obtain the multimodal approach, we applied score level fusion with different combinations of the monomodal attention level estimation classifiers, therefore, we sorted out our systems into unimodal and multimodal attention level estimation. The training process works as follows:

1) UNIMODAL ATTENTION LEVEL ESTIMATION

i) Each frame is processed through the 5 facial analysis modules described in section III-B. *ii)* Output features $\mathbf{f}_{x,y}$ are averaged for each second of video $\bar{\mathbf{f}}_{x,y}$. *iii)* A vector $\mathbf{f}_{\mathrm{L}}^{y}$ for local and $\mathbf{f}_{\mathrm{G}}^{y}$ for global features are obtained for the time window at hand. The extraction process follows the steps described in the previous section III-C. Finally, we have the following vectors $\{\mathbf{f}_{\mathrm{L}}^{\mathrm{EB}}, \mathbf{f}_{\mathrm{L}}^{\mathrm{EAR}}, \mathbf{f}_{\mathrm{L}}^{\mathrm{NS}}, \mathbf{f}_{\mathrm{L}}^{\mathrm{H}}, \mathbf{f}_{\mathrm{L}}^{\mathrm{HS}}, \mathbf{f}_{\mathrm{L}}^{\mathrm{HP}}, \mathbf{f}_{\mathrm{L}}^{\mathrm{Exp}}\}$ and $\{\mathbf{f}_{\mathrm{G}}^{\mathrm{EB}}, \mathbf{f}_{\mathrm{G}}^{\mathrm{EAR}}, \mathbf{f}_{\mathrm{G}}^{\mathrm{HS}}, \mathbf{f}_{\mathrm{G}}^{\mathrm{NS}}, \mathbf{f}_{\mathrm{G}}^{\mathrm{H}}, \mathbf{f}_{\mathrm{G}}^{\mathrm{HP}}, \mathbf{f}_{\mathrm{G}}^{\mathrm{Exp}}\}$. *iv)* Two SVMs for each facial feature category are trained to classify between high and low attention, one using local features $\mathbf{f}_{\mathrm{L}}^{y}$ and the other using global features $\mathbf{f}_{\mathrm{G}}^{y}$ as input. The scores for local features are denoted as s_{L}^{y}, which include $\{s_{\mathrm{L}}^{\mathrm{EB}}, s_{\mathrm{L}}^{\mathrm{EAR}}, s_{\mathrm{L}}^{\mathrm{H}}, s_{\mathrm{L}}^{\mathrm{NS}}, s_{\mathrm{L}}^{\mathrm{HS}}, s_{\mathrm{L}}^{\mathrm{HP}}, s_{\mathrm{L}}^{\mathrm{Exp}}\}$ and for global features as s_{G}^{y}, which include $\{s_{\mathrm{G}}^{\mathrm{EB}}, s_{\mathrm{G}}^{\mathrm{EAR}}, s_{\mathrm{G}}^{\mathrm{NS}}, s_{\mathrm{G}}^{\mathrm{HS}}, s_{\mathrm{G}}^{\mathrm{HP}}, s_{\mathrm{G}}^{\mathrm{H}}, s_{\mathrm{G}}^{\mathrm{Exp}}\}$.

2) MULTIMODAL ATTENTION LEVEL ESTIMATION

The proposed multimodal systems involve combining unimodal facial analysis systems based on either local or global features. The scores from previously trained unimodal facial analysis are combined using different strategies: *i)* a score sum strategy and *ii)* training a simple neural network with two hidden layers. The architecture consists of dense layers with ReLU activation. The first hidden layer has 16 units and processes the input, which includes 7 scores, each corresponding to the output from the SVM binary classifiers for individual facial feature categories. This is followed by another dense layer with 8 units, and an output layer with one unit (sigmoid activation). A dropout of 0.5 is employed. For both fusion strategies, the process was carried out individually for local features s_{L}^{y} and for global features s_{G}^{y}, obtaining two combined scores $s_{\mathrm{L}}^{\mathrm{F}}$ or $s_{\mathrm{G}}^{\mathrm{F}}$. Finally, these scores were compared with a threshold τ to determine the attention level (high or low).

We also propose another multimodal system for global features, based on feature selection and fusion using a single SVM classifier (see Fig. 5). The protocol is the same as previously explained; however, instead of training an SVM for each facial feature category, we perform a feature selection and fusion inspired by the work of Leng et al. [44]. We merged global features into a vector $\{\mathbf{f}_{\mathrm{G}}^{\mathrm{EB}}, \mathbf{f}_{\mathrm{G}}^{\mathrm{EAR}}, \mathbf{f}_{\mathrm{G}}^{\mathrm{HS}}, \mathbf{f}_{\mathrm{G}}^{\mathrm{H}}, \mathbf{f}_{\mathrm{G}}^{\mathrm{NS}}, \mathbf{f}_{\mathrm{G}}^{\mathrm{HP}}, \mathbf{f}_{\mathrm{G}}^{\mathrm{Exp}}\}$ and calculate the Discrimination Power (DP), which is a measure based on the inter-class and intra-class variation $\mathrm{DP} = \frac{\sigma_{\mathrm{inter}}^{2}}{\sigma_{\mathrm{intra}}^{2}}$. Finally, we select the features that are in the top 90th percentile of DP. This new vector $\mathbf{f}_{\mathrm{G}}^{\mathrm{S}}$ is used as input to train a single SVM to classify between high and low attention.

IV. EXPERIMENTS AND RESULTS

A. EXPERIMENTAL PROTOCOL

We follow the protocol proposed in ALEBK [11] to classify between high and low attention levels, as detailed in the previous section III-D. In total, we obtain 10376, 8309, and 5605 periods for time windows W_l of 30, 60, and 120 seconds, respectively, from all 60 students in the database. The samples are evenly distributed between low and high attention levels.

We employ the leave-one-out cross-validation protocol, where one user is left out for testing, and the remaining ones are used for training and this process is repeated with all users. The decision threshold is chosen at the point where the classification accuracy is maximized.

TABLE 4. Attention estimation Accuracy (Acc in %) using the mEBAL2 database for the proposed unimodal approaches with local features. We set the value of τ_L at 10% and τ_H at 90%. The values highlighted in black indicate the best module for each time window (30s, 60s, 120s).

Module	W_l : 30 Seconds Acc	W_l : 60 Seconds Acc	W_l : 120 Seconds Acc
Landmark (EAR)	68.52	69.84	75.54
EyeBlink (EB)	70.54	73.91	**79.16**
Expression (Exp)	**76.66**	**77.28**	79.11
Head Pose (HP)	57.21	60.61	65.23
Landmark (HS)	61.66	62.78	65.94
Landmark (NS)	62.33	62.20	57.22
Heart Rate (H)	50.03	50.02	54.83

W_l: Window length (in seconds).

TABLE 5. Attention estimation Accuracy (Acc in %) using the mEBAL2 database for the proposed unimodal approaches with global features. We set the value of τ_L at 10% and τ_H at 90%. The values highlighted in black indicate the best module for each time window (30s, 60s, 120s).

Module	W_l: 30 Seconds Acc	W_l: 60 Seconds Acc	W_l: 120 Seconds Acc
Landmark (EAR)	**75.94**	**75.87**	75.99
EyeBlink (EB)	73.03	74.41	**80.64**
Expression (Exp)	73.05	74.46	78.39
Head Pose (HP)	63.62	59.78	59.52
Landmark (HS)	64.12	62.09	53.79
Landmark (NS)	63.07	62.17	56.32
Heart Rate (H)	52.35	55.84	59.86

W_l: Window length (in seconds).

B. UNIMODAL EXPERIMENTS

We initially divided the experiments into local and global features.

1) LOCAL FEATURES

Table 4 displays the results for each facial analysis module in terms of attention estimation Accuracy (Acc in %) for all time windows (30s, 60s, 120s). Fig. 6 shows the probability density distributions of the scores obtained for the best method in each window.

The results show that the EyeBlink (EB) and Facial Expression (Exp) modules achieve the highest accuracy with better separability between distributions for all time frames. We noticed that in the 30s and 60s windows, the Exp module performs the best with an accuracy of 76.66% and 77.28%, respectively. However, in the 120s window, the EB module shows a slight improvement over Exp, achieving an accuracy of 79.16%. The third module (feature category) with the best results is the EAR feature category, which reinforces previous findings on the importance of the eye state and facial expressions in attention estimation [11], [17], [38], [39].

The worst results are obtained from the Heart Rate (H) module. This suggests that, the variations in Heart Rate do not present a high correlation with attention levels in this database.

The Head Pose (HP) module has the second worst result for 30s and 60s windows; however, even though it is not a clear attention estimation indicator, it shows that there is a relationship with attention levels, making it potentially useful for multimodal approaches. Additionally, it is observed that as the time window increases, the results improve, reaching an accuracy of 65.23%. This makes sense, as a larger window allows capturing more significant patterns and trends in the student's behavior and mitigates possible errors from the pose detection module.

The previous modules show an improvement in the accuracy metric when the time window is extended, with an average improvement of around 6.18%. This demonstrates that increasing the amount of features and context allows a better classification. The feature categories based on the eye state are particularly relevant, specifically the EB and EAR, where we observe an accuracy improvement of 8.62% and 7.02% respectively. This makes sense because eyeblinks are less frequent in e-learning environments compared to standard behavior [71], [72]. For this reason, a larger window allows the detection of moments with few eyeblinks (high attention) or periods with a higher eyeblink frequency (low attention).

Similar to HP, head-to-camera indicators like Head and Nose Size, HS and NS respectively, are not strongly correlated to attention. Unlike previous modules, these feature categories do not always perform better in the 120-second window. This makes sense because during e-learning sessions, students can make fast movements to get closer to the screen, fixing their visual attention on a specific point on the screen, indicating strong concentration.

Fig. 6 shows that, in most cases, high attention levels are easier to recognize than the low ones. Low levels tend to have a more spread density distribution, making their classification more challenging. This makes sense in the context of the monitoring carried out in mEBAL2 [16], where students are typically focused with moments of high attention during short time tasks.

2) GLOBAL FEATURES

We conducted the same experiments as in the previous section with the global features to understand if they are more effective in the SVM-based classification and how they impact each module. This analysis aims to assess whether the global features can provide additional discriminating information to improve the accuracy of attention estimation compared to the local features.

Table 5 shows the results for each module in different time windows (30s, 60s, 120s). Similar to the previous case, the probability density distributions of the scores for the best method in each window are shown in Fig. 7.

The EAR feature category achieves the best results in the 30s and 60s windows, achieving a maximum accuracy of 75.94% and 75.87%, respectively. We can observe significant improvements in the results of this module in comparison to local features, achieving an accuracy improvement of 7.42% and 6.03%, respectively. Once again, the top three feature categories with the best results are EAR, EB, and Exp.

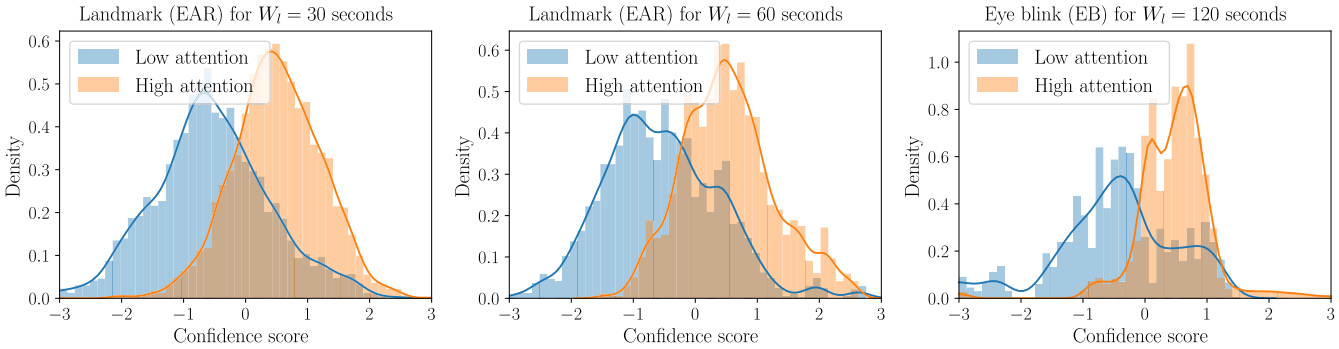


FIGURE 7. Probability density distributions of the confidence scores obtained by our attention estimation systems for the best approach in each of the three time windows considered (from left to right: 30s, 60s, and 120s) using global features. In order to simplify the performance analysis/comparison, our experimental discussion is focused on binary classification into low/high attention using the score threshold that maximizes classification accuracy.

TABLE 6. Accuracy results (Acc in %) for attention estimation in multimodal systems based on local features, showing the best combinations for score sum fusion. The first row provides the best unimodal module for the selected time window. The last row displays the results achieved by score fusion via neural network. The values highlighted in black indicate the best feature categories and the fusion strategy with the best accuracy for each time window (30s, 60s, 120s).

W_l: 30 Seconds		W_l: 60 Seconds		W_l: 120 Seconds	
Feature Categories	**Acc**	**Feature Categories**	**Acc**	**Feature Categories**	**Acc**
Exp	76.66	Exp	77.28	EB	79.16
EB, Exp	**77.25**	EB, Exp	**77.65**	EB, Exp	**80.52**
EB, Exp, HS	73.95	EB, Exp, HP	75.92	EB, Exp, HP	80.32
Exp, EAR, HP, HS	73.00	EB, Exp, EAR, HP	76.48	EB, Exp, EAR, H	79.95
EB, Exp, EAR, HP, HS	73.18	EB, Exp, EAR, HP, HS	75.11	EB, Exp, EAR, HP, H	78.18
EB, Exp, EAR, HP, H, HS	70.46	EB, Exp, EAR, HP, HS, NS	72.10	EB, Exp, EAR, HP, H, HS	76.88
All Modules	68.67	All Modules	70.01	All Modules	76.25
Neural Network Fusion	**84.25**	Neural Network Fusion	**85.87**	Neural Network Fusion	**85.92**

W_l: Window length (in seconds).

In the case of the EB module, we can see improvements in all three windows, but a notable difference in the 120-second window. The accuracy in this case reaches 80.64%, which is the highest obtained value.

The Exp module shows a decrease in accuracy results compared to the local features in all three windows, with differences of 3.61% for the 30s window, 2.82% for the 60s window, and 0.72% for the 120s window, noticing an error reduction as the window size increases.

The Heart Rate module remains an unreliable indicator for attention estimation, as its classification is almost random in the considered time windows. The other features categories, user distance and head pose, exhibit similar behavior, showing slight improvements in the first window and deterioration in the subsequent ones when compared to local features. By themselves do not serve as a clear indicator for attention estimation. However, as we will see later, the information provided by these features categories might be valuable in multimodal systems.

Results show that our best unimodal models improve their performance as the temporal window increases up to 120 seconds. The same trend is observed with local features, highlighting the importance of considering a longer time period to capture significant patterns and trends in the students' behavior. This finding supports the notion that certain discriminating features may become clearer and more effective in attention estimation when analyzing a larger temporal context. By expanding the window, we allow the modules to detect and utilize more relevant information for classification, resulting in an enhanced ability to distinguish between high and low attention levels with greater accuracy. As we can see, some modules were significantly improved using global features, such as EAR feature category. Additionally, the size of the temporal windows has a notable impact on the results. Global features achieve the highest accuracy value of 80.64% for the EB module.

Figure 7 also shows that detecting low attention levels can be more challenging than detecting higher ones, because the low attention score distribution is more spread than the high attention one. Although this difference is not as clear in global features as it is in local ones, it is particularly evident in the 120s window.

C. MULTIMODAL EXPERIMENTS

1) LOCAL FEATURES

Table 6 displays the results from the best combinations of unimodal for the score sum strategy and the score fusion results using a neural network.

TABLE 7. Accuracy results (Acc in %) for attention estimation in multimodal systems based on global features, showing the best combinations for score sum fusion. The first row provides the best unimodal module for the selected time window. The last row displays the results achieved by score fusion via neural network. The values highlighted in black indicate the best feature categories and the fusion strategy with the best accuracy for each time window (30s, 60s, 120s).

W_l: 30 Seconds		W_l: 60 Seconds		W_l: 120 Seconds	
Feature Categories	**Acc**	**Feature Categories**	**Acc**	**Feature Categories**	**Acc**
EAR	75.94	EAR	75.87	EB	80.64
EAR, Exp	76.77	EB, EAR	79.17	EB, Exp	**83.34**
EB, EAR, HS	77.11	EB, EAR, NS	78.90	EB, Exp, EAR	80.95
EB, Exp, EAR, HP	77.35	EB, Exp, EAR, NS	79.01	EB, Exp, EAR, NS	79.80
EB, Exp, EAR, HP, HS	**77.39**	EB, Exp, EAR, HP, NS	**79.23**	EB, Exp, EAR, H, NS	77.23
EB, Exp, EAR, HP, HS, H	77.31	EB, Exp, EAR, HP, H, NS	78.32	EB, Exp, EAR, H, HS, NS	76.16
All Modules	76.73	All Modules	77.05	All Modules	73.10
Neural Network Fusion	**79.26**	Neural Network Fusion	79.17	Neural Network Fusion	74.01

W_l: Window length (in seconds).

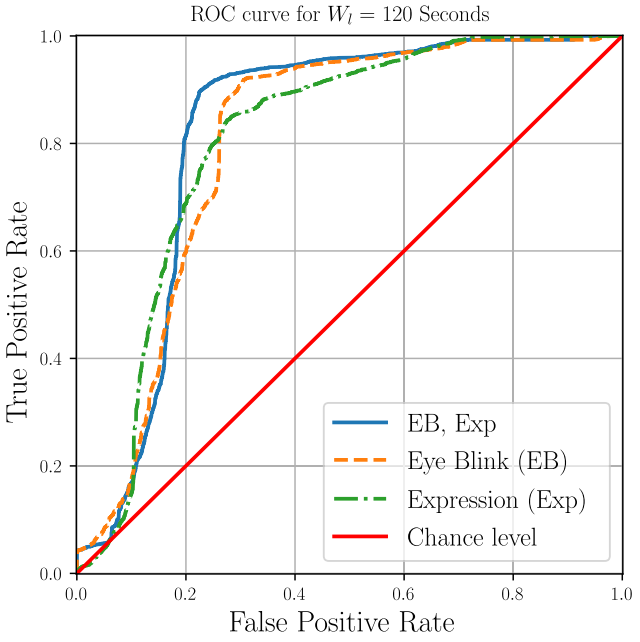


FIGURE 8. Receiver Operating Characteristic curve (ROC) obtained for the most accurate multimodal approach using global features (this occurs in the 120s window), shown with a blue line and for each of the monomodal systems that are part of this combination.

The best results in the 30s window for score sum are achieved combining the EB+Exp modules, with an accuracy of 77.25%. Compared to the Exp module, which is the best unimodal module, there is a slight improvement of 0.60%. We observe that the combination with other modules worsens the results compared to the Exp module. However, the score fusion using a neural network achieves the best performance with 84.25%, marking a significant improvement over the EB+Exp module combination by 7%, demonstrating the potential of neural networks for score fusion [43].

The same pattern occurs in the 60s window, making the EB+Exp combination the best for score sum, showing a slight improvement over the Exp module alone. The other combinations result in a worse performance. Once again, in the 60s window, the Neural Network Fusion (NNF) achieves the best results, even surpassing those in the 30s window. NNF outperforms both the unimodal system, with a significant improvement of 8.59%, and the top-performing EB+Exp combination by 8.22%.

In the 120s window the best unimodal system is EB, which is slightly surpassed by three different fusions: EB+Exp, EB+Exp+HP, and EB+Exp+EAR+H. The best maximum accuracy is achieved by EB+Exp with 80.52% for score sum. Once again, NNF outperforms the score sum, achieving an accuracy of 85.92%. This demonstrates that increasing the temporal window improves the system combination accuracy, as expected because a broader temporal context facilitates the integration of longer and more complex temporal patterns in the data, resulting in better discrimination. Furthermore, these results show that score-level fusion with neural networks is more effective than score sum for local features in the mEBAL2 database.

The EB and Exp unimodal modules are the most effective in attention estimation, appearing in all combinations that improved results. Additionally, the best values are consistently obtained in the 120s window, having the most potential for improvement due to the wider amount of information. This highlights the challenge of attention estimation through image processing, requiring longer windows to capture relevant behavioral and physiological processes.

2) GLOBAL FEATURES

Table 7 presents the results of the best combinations with global features for score sum and neural network fusion. Fig. 8 shows the ROC curve for the best multimodal system, along with the results of the individual monomodal systems that compose it.

With global features, more effective combinations are achieved for score sum and lower performances are obtained for NNF. In the 30s window, the best unimodal result is 75.94%, and all combinations shown in the table (combining with the EAR feature category) outperform it,

including the combination of all modules. The best one is EB+Exp+EAR+HP+HS with 77.39%, representing an average improvement of 1.65% in accuracy. Furthermore, this combination slightly outperforms (0.14%) the best result with local features, which was EB+Exp in the 30s window. However, EB+Exp requires only two modules, making it faster and more practical. NNF achieves better performance than the best combination by score sum, with a slight improvement of 1.87%. However, its performance is inferior to the results obtained for local features in all windows.

In the 60s window, we observe a similar pattern for the score sum. The unimodal EAR feature category achieves an accuracy of 75.87% and all the combinations outperform it when combined with the same feature category. The best combination is EB+Exp+EAR+HP+NS, similar to the previous one, with an accuracy of 79.23%. This combination shows a significant improvement of 3.36% in accuracy. This highlights that the user distance and pose feature categories contain valuable information in multimodal systems, especially in short-duration windows. Furthermore, the best combination of global features surpasses the results of the best combination with local features, EB+Exp, by 1.58% in accuracy, demonstrating a considerable improvement. However, for NNF, the results are similar to those in the 30s window. For the first time, this method is inferior to the best combination for score sum, though almost equal. Once again, the results obtained by NNF for global features are inferior to those achieved with local features.

For the 120s window, we found out that the combination of EB+Exp significantly outperformed the best unimodal approach, which was EB. On the other hand, EB+Exp achieves 83.34% resulting in a remarkable improvement of 2.7% in accuracy. Furthermore, global features also surpass the best combination with local features for score sum by 2.82%, showing that global features continue to achieve a better combination of modules, especially in the 120s window where the best results are obtained. Additionally, in this window, the best combinations for both local and global features are EB+Exp, indicating that, under similar conditions, the most effective system is achieved with global features. Additionally, this also highlights the importance of facial units and EyeBlink in attention estimation. NNF achieves its lowest performance with an accuracy of 74.01%, which is 9.33% lower than the best combination of EB+Exp. It has been observed that score fusion using global features has worse generalization compared to local features

Figure 8 presents the ROC curve for the unimodal and multimodal approaches based on global features and 120s window (best approaches). The curve shows significant improvement when combining Facial Expressions and EyeBlink.

3) GLOBAL FEATURES: SELECTION AND FEATURE LEVEL FUSION

Fig. 5 shows the proposed architecture for feature selection and fusion (for more details, see Section III-D). Table 8 presents the best results achieved for unimodal systems, score level fusion, and feature level fusion using global features.

TABLE 8. Best results of accuracy for global features in unimodal system, score level fusion, and feature level fusion with the best accuracy for each time window (30s, 60s, 120s).

Methods	W_l: 30s Acc	W_l: 60s Acc	W_l: 120s Acc
Unimodal System	75.94	75.87	80.64
Score Level Fusion	79.26	79.23	83.34
Feature Level Fusion	76.48	77.79	81.48

TABLE 9. Comparison with the state of the art. Attention level estimation results on the mEBAL2 dataset [16] including 60 students. Our best approach is compared with Peng [13], ALEBk [11] and MATT [12]. The same training and evaluation protocol is employed for all methods following our experimental protocol. *We have adapted the method Peng [13] for classifying between high and low attention levels. This method was designed to work with global features extracted from the head pose module and the facial landmark module. **We have adapted the methods proposed in [11] and [12] incorporating the global and local features proposed in this work. The results obtained for the 120s time frame are shown, which exhibited the highest accuracy for the best-performing approaches.

Methods	Local Features Acc	Global Features Acc
Peng [40]*	–	66.28
ALEBk** [11]	79.16	80.64
MATT** [12]	80.32	74.43
Proposed: DeepFace-Attention	**85.92**	83.34

The results show that the proposed architecture for feature selection and fusion outperforms unimodal systems in all windows. However, it produces inferior results compared to the top-performing multimodal systems achieved through score fusion. However, our feature level fusion architecture only utilizes 10% of the global features (reducing from 728 features to 73) and uses only an SVM to obtain a direct score without the need to train a neural network, which requires more careful optimization. This demonstrates that the results of this architecture are highly competitive.

D. EXPERIMENTS: COMPARISON WITH EXISTING APPROACHES

We now compare ourselves with three recent state-of-the-art approaches: Peng [13], ALEBk [11], and MATT [12]. ALEBk and MATT were previously trained and evaluated on the first version of mEBAL [17]. To perform the comparison here, we train again all the methods on mEBAL2 with 60 users under identical conditions, using the same attention classification percentile, and employing the leave-one-out cross-validation protocol. Table 9 presents a benchmark with the best results obtained in attention estimation on the mEBAL2 dataset [16] by different state-of-the-art approaches, compared to our proposal here: DeepFace-Attention.

The approach proposed in Peng et al. [13] is a multimodal system based on global features of head posture and movements of the eyes, head, and mouth (see Section II-B for more information). This approach was based on a random forest model to estimate attention in 10s windows. We adapted this model to predict high and low attention in 30s, 60s, and 120s windows. The results obtained are inferior compared to the methods ALEBk [11], MATT [12], and our method. Regarding global features, our method improves the performance by 17.06% over Peng et al. [13] and by 19.64% over our best method based on local features.

ALEBk [11] is a monomodal system that estimates attention based on the eyeblink rate per minute. An enhanced version of that system was used, incorporating an SVM for high and low attention classification, using local and global features, obtained from the eyeblink detector, rather than simply applying a blink rate per minute threshold. While ALEBk achieved an accuracy of 74% for the first version of mEBAL, the improved version obtains 79.16% when applied to mEBAL2 with 60 users. Furthermore, the results of employing global features have been also evaluated over ALEBk, achieving an accuracy of 80.64%.

MATT [12] presented unimodal and multimodal approaches to classify between high and low attention levels. Its best-performing approach was the multimodal one, which combined EyeBlink, Head Pose, and Facial Expression, using local features. This method achieved an accuracy of 80.32% with local features and 74.43% with global features. Better results are obtained with local features compared to the global ones, in contrast to the outcomes achieved by ALEBk.

As seen on Table 9, our approach outperforms previous approaches. Our best multimodal approach is EB+Exp score sum combination for global features and NNF for local features. The best results are achieved with local features, surpassing the latest version of ALEBk by 6.8%, resulting in a relative reduction in error rates of 32.4%. Regarding the best version of MATT, corresponding to the use of local features, an improvement of 5.6% in accuracy is obtained, leading to a relative reduction in error rates of 28.5%. Our global feature system based on score sum also outperforms state-of-the-art proposals using global features with a relative reduction in error rates of 14% for ALEBk and 34.8% for MATT.

Furthermore, our multimodal system based on score sum requires only two modules (EB+Exp), while the MATT approach requires three (EB+Exp+HP), resulting in reduced time and resource usage.

Table 10 presents the results of the average inference speed for each processed frame by different facial modules using an Intel Core i5-7600 CPU, 32GB of RAM, and a NVIDIA GTX 1080 GPU with 8GB of VRAM, providing a computational comparison of the various modules. It also presents the inference times per processed frame for our methods and the state-of-the-art methods: Peng et al. [13], ALEBk [11], and MATT [12]. The Head Pose module is the slowest, followed by the landmark module. Although the EyeBlink module only takes 15.94 ms, using the landmark module is necessary to identify the eye region, thus the total time of 57 ms. As we can see in Table 10, the slowest methods are the systems that utilize the Head Pose, such as Peng et al. [13], MATT [12], and our local features NNF method. The fastest method is ALEBk [11], which only uses the eyeblink and landmark modules.

TABLE 10. Comparative Inference Times for Attention Estimation using an Intel Core i5-7600 CPU, RAM 32GB of RAM, and a NVIDIA GTX 1080 GPU with 8GB of VRAM. This table includes inference times of the facial analysis modules and a comparison with state-of-the-art methods. Our best approach is compared with Peng et al. [13], ALEBk [11], and MATT [12].

Modules	Inference speed (ms)
Face Detection	29.76
EyeBlink	15.94
Landmark	41.45
Head Pose	101.57
Expression	14.59
Heart Rate	5.15
Methods	**Inference speed (ms)**
Peng [13]	172
ALEBk [11]	86
MATT [12]	202
Proposed (Global Features)	100
Proposed (Local Features)	208

Note that our objective in this research was not to enhance the system's speed but rather to evaluate whether deep learning-based facial analysis modules can accurately determine high or low attention levels. For future work, resource-optimized modules can be used to reach real-time operation if needed.

V. CONCLUSION

We have presented various approaches to estimate high or low attention levels, applied to a realistic e-learning environment of 60 students. State-of-the-art technologies were used, based on deep learning, to perform facial analysis of behavioral features and physiological processes related to attention [12], [14], [16]. To understand which features are more efficient in attention estimation, we designed unimodal systems based on SVM classification using the following information: eyeblink, heart rate, facial expressions, head pose, and head distance. We also have investigated the impact of local features and well-known global features on accuracy. Additionally, we examined the effects of temporal windows on attention estimation, with three different options: 30, 60, and 120 seconds. We proposed multimodal systems for attention estimation, demonstrating their ability to enhance existing methods for attention estimation.

Some interesting findings are as follows: eye state features (EAR, EyeBlink) and facial expressions are the most useful with a clear correlation with attention. We also observed

that the best attention estimation systems improved as the time window size increases. Head pose and distance features were not clear indicators of attention; however, in multimodal systems, they provided relevant information for classification. The results of the Heart Rate module, both unimodal and combined, showed that it is not a reliable indicator of attention. Global features were more effective for multimodal systems based on score sum, obtaining the best combination with Eyeblink and Facial Expressions with an accuracy of 83.34%. The best results in this study were achieved with local features using score level fusion through neural network training with an accuracy of 85.92%. We also analyzed an architecture based on the selection and fusion of global features, outperforming unimodal systems with slightly less accuracy than our full score fusion, but only necessitating 10% of the features.

Our best approach, called DeepFace-Attention, have outperformed three state-of-the-art methods: Peng et al. [13], ALEBk [11], and MATT [12]; achieving a significant relative improvement in error reduction of approximately 50.6% for Peng et al. [13], 32.4% for ALEBk (an enhanced version of the system proposed in [11]), and 28.5% for MATT.

In the future, we will explore the combination of local and global features during the training process. Moreover, we aim to analyze how attention estimation can be affected when students perform different types of tasks. Additionally, we will explore alternative indicators that have shown a direct relation with attention levels, such as eye pupil size [25], [26], gaze tracking [31], [73], keystroking [74], [75], [76], among others. Predicting the level of attention within a continuous range is a more challenging task than predicting high or low attention levels, and it is also planned for future work.

REFERENCES

[1] H. Tang, M. Dai, S. Yang, X. Du, J.-L. Hung, and H. Li, "Using multimodal analytics to systemically investigate online collaborative problem-solving," *Distance Educ.*, vol. 43, no. 2, pp. 290–317, Apr. 2022.

[2] D. Kahneman, Ed., *Attention and Effort*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1973.

[3] L. M. Bergasa, J. Nuevo, M. A. Sotelo, R. Barea, and M. E. Lopez, "Real-time system for monitoring driver vigilance," *IEEE Trans. Intell. Transp. Syst.*, vol. 7, no. 1, pp. 63–77, Mar. 2006.

[4] J. Hernandez-Ortega, S. Nagae, J. Fierrez, and A. Morales, "Quality-based pulse estimation from NIR face video with application to driver monitoring," in *Proc. Iberian Conf. Pattern Recognit. Image Anal.*, 2019, pp. 108–119.

[5] M. Wedel and R. Pieters, "Eye tracking for visual marketing," *Found. Trends Marketing*, vol. 1, no. 4, pp. 231–320, 2006.

[6] C. G. Lim, T. S. Lee, C. Guan, D. S. S. Fung, Y. Zhao, S. S. W. Teng, H. Zhang, and K. R. R. Krishnan, "A brain-computer interface based attention training program for treating attention deficit hyperactivity disorder," *PLoS ONE*, vol. 7, no. 10, Oct. 2012, Art. no. e46692.

[7] S. Leal and A. Vrij, "Blinking during and after lying," *J. Nonverbal Behav.*, vol. 32, no. 4, pp. 187–194, Dec. 2008.

[8] S. Mann, A. Vrij, and R. Bull, "Suspects, lies, and videotape: An analysis of authentic high-stake liars," *Law Hum. Behav.*, vol. 26, no. 3, pp. 365–376, Jun. 2002.

[9] S. T. Iqbal, X. S. Zheng, and B. P. Bailey, "Task-evoked pupillary response to mental workload in human-computer interaction," in *Proc. CHI Extended Abstr. Human Factors Comput. Syst.*, Apr. 2004, pp. 1477–1480.

[10] J. Hernandez-Ortega, R. Daza, A. Morales, J. Fierrez, and J. Ortega-Garcia, "edBB: Biometrics and behavior for assessing remote education," in *Proc. AAAI Workshop Artif. Intell. Educ.*, 2020, pp. 1–7.

[11] R. Daza, D. DeAlcala, A. Morales, R. Tolosana, R. Cobos, and J. Fierrez, "ALEBk: Feasibility study of attention level estimation via blink detection applied to e-learning," in *Proc. AAAI Workshop Artif. Intell. Educ.*, 2022, pp. 1–7.

[12] R. Daza, L. F. Gomez, A. Morales, J. Fierrez, R. Tolosana, R. Cobos, and J. Ortega-Garcia, "MATT: Multimodal attention level estimation for e-learning platforms," in *Proc. AAAI Workshop Artif. Intell. Educ.*, 2023, pp. 1–7.

[13] S. Peng, L. Chen, C. Gao, and R. J. Tong, "Predicting students' attention level with interpretable facial and head dynamic features in an online tutoring system (student abstract)," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 13895–13896.

[14] J. Zaletelj and A. Košir, "Predicting students' attention in the classroom from Kinect facial and body features," *EURASIP J. Image Video Process.*, vol. 2017, no. 1, pp. 1–12, Dec. 2017.

[15] J. Fierrez, A. Morales, R. Vera-Rodriguez, and D. Camacho, "Multiple classifiers in biometrics. Part 1: Fundamentals and review," *Inf. Fusion*, vol. 44, pp. 57–64, Nov. 2018.

[16] R. Daza, A. Morales, J. Fierrez, R. Tolosana, and R. Vera-Rodriguez, "MEBAL2 database and benchmark: Image-based multispectral eyeblink detection," *Pattern Recognit. Lett.*, vol. 182, pp. 83–89, Jun. 2024.

[17] R. Daza, A. Morales, J. Fierrez, and R. Tolosana, "MEBAL: A multimodal database for eye blink detection and attention level estimation," in *Proc. Companion Publication Int. Conf. Multimodal Interact.*, Oct. 2020, pp. 32–36.

[18] J. E. Hall and M. E. Hall, *Guyton and Hall Textbook of Medical Physiology E-Book*. Amsterdam, The Netherlands: Elsevier, 2020.

[19] T. Kirschstein and R. Köhling, "What is the source of the EEG?" *Clin. EEG Neurosci.*, vol. 40, no. 3, pp. 146–149, Jul. 2009.

[20] X. Li, B. Hu, T. Zhu, J. Yan, and F. Zheng, "Towards affective learning with an EEG feedback approach," in *Proc. 1st ACM Int. Workshop Multimedia Technol. Distance Learn.*, Oct. 2009, pp. 33–38.

[21] C.-M. Chen and J.-Y. Wang, "Effects of online synchronous instruction with an attention monitoring and alarm mechanism on sustained attention and learning performance," *Interact. Learn. Environ.*, vol. 26, no. 4, pp. 427–443, May 2018.

[22] Y. Li, X. Li, M. Ratcliffe, L. Liu, Y. Qi, and Q. Liu, "A real-time EEG-based BCI system for attention recognition in ubiquitous environment," in *Proc. Int. Workshop Ubiquitous Affect. Awareness Intell. Interact.*, Sep. 2011, pp. 33–40.

[23] E. Haapalainen, S. Kim, J. F. Forlizzi, and A. K. Dey, "Psycho-physiological measures for assessing cognitive load," in *Proc. 12th ACM Int. Conf. Ubiquitous Comput.*, Sep. 2010, pp. 301–310.

[24] J. Hernandez-Ortega, R. Daza, A. Morales, J. Fierrez, and R. Tolosana, "Heart rate estimation from face videos for student assessment: Experiments on edBB," in *Proc. IEEE 44th Annu. Comput., Softw., Appl. Conf. (COMPSAC)*, Jul. 2020, pp. 172–177.

[25] S. Rafiqi, C. Wangwiwattana, J. Kim, E. Fernandez, S. Nair, and E. C. Larson, "PupilWare: Towards pervasive cognitive load measurement using commodity devices," in *Proc. 8th ACM Int. Conf. Pervasive Technol. Rel. Assistive Environments*, Jul. 2015, pp. 1–8.

[26] K. Krejtz, A. T. Duchowski, A. Niedzielska, C. Biele, and I. Krejtz, "Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze," *PLoS ONE*, vol. 13, no. 9, Sep. 2018, Art. no. e0203629.

[27] W. Boucsein, *Electrodermal Activity*. Berlin, Germany: Springer, 2012.

[28] Z. Luo, C. Jingying, W. Guangshuai, and L. Mengyi, "A three-dimensional model of student interest during learning using multimodal fusion with natural sensing technology," *Interact. Learn. Environments*, vol. 30, no. 6, pp. 1117–1130, Jul. 2022.

[29] M. Raca, L. Kidzinski, and P. Dillenbourg, "Translating head motion into attention-towards processing of student's body-language," in *Proc. 8th Int. Conf. Educ. Data Mining*, 2015, pp. 1–7.

[30] A. Becerra, J. Irigoyen, R. Daza, R. Cobos, A. Morales, J. Fierrez, and M. Cukurova, "Biometrics and behavior analysis for detecting distractions in e-learning," in *Proc. Int. Symp. Comput. Educ. (SIIE)*, 2024, pp. 1–6.

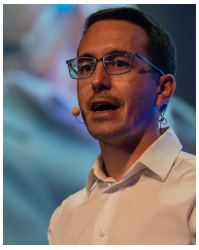
[31] Q. Wang, S. Yang, M. Liu, Z. Cao, and Q. Ma, "An eye-tracking study of website complexity from cognitive load perspective," *Decis. Support Syst.*, vol. 62, pp. 1–10, Jun. 2014.

[32] Á. Becerra, R. Daza, R. Cobos, A. Morales, M. Cukurova, and J. Fierrez, "M2LADS: A system for generating MultiModal learning analytics dashboards," in *Proc. IEEE 47th Annu. Comput., Softw., Appl. Conf. (COMPSAC)*, Jun. 2023, pp. 1564–1569.
[33] H. Monkaresi, N. Bosch, R. A. Calvo, and S. K. D'Mello, "Automated detection of engagement using video-based estimation of facial expressions and heart rate," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 15–28, Jan. 2017.
[34] B. McDaniel, S. D'Mello, B. King, P. Chipman, K. Tapp, and A. Graesser, "Facial features for affective state detection in learning environments," in *Proc. Annu. Meeting Cognit. Sci. Soc.*, 2007, pp. 1–7.
[35] J. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, and J. Lester, "Automatically recognizing facial expression: Predicting engagement and frustration," in *Proc. Educ. Data Mining*, 2013, pp. 1–8.
[36] K. Fujisawa and K. Aihara, "Estimation of user interest from face approaches captured by webcam," in *Proc. Intl. Conf. Virtual Mixed Reality*, 2009, pp. 51–59.
[37] L. Fridman, B. Reimer, B. Mehler, and W. T. Freeman, "Cognitive load estimation in the wild," in *Proc. CHI Conf. Human Factors Comput. Syst.*, Apr. 2018, pp. 1–9.
[38] J. Bagley and L. Manelis, "Effect of awareness on an indicator of cognitive load," *Perceptual Motor Skills*, vol. 49, no. 2, pp. 591–594, Dec. 1979.
[39] M. K. Holland and G. Tarlow, "Blinking and mental load," *Psychol. Rep.*, vol. 31, no. 1, pp. 119–127, Aug. 1972.
[40] T. Soukupovà and J. Cech, "Real-time eye blink detection using facial landmarks," in *Proc. Comput. Vis. Winter Workshop*, 2016, pp. 1–8.
[41] P. Goldberg, Ö. Sümer, K. Stürmer, W. Wagner, R. Göllner, P. Gerjets, E. Kasneci, and U. Trautwein, "Attentive or not? Toward a machine learning approach to assessing students' visible engagement in classroom instruction," *Educ. Psychol. Rev.*, vol. 33, no. 1, pp. 27–49, Mar. 2021.
[42] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 59–66.
[43] J. Fierrez, A. Morales, R. Vera-Rodriguez, and D. Camacho, "Multiple classifiers in biometrics. Part 2: Trends and challenges," *Inf. Fusion*, vol. 44, pp. 103–112, Nov. 2018.
[44] L. Leng, M. Li, C. Kim, and X. Bi, "Dual-source discrimination power analysis for multi-instance contactless palmprint recognition," *Multimedia Tools Appl.*, vol. 76, no. 1, pp. 333–354, Jan. 2017.
[45] L. Leng and J. Zhang, "PalmHash code vs. PalmPhasor code," *Neurocomputing*, vol. 108, pp. 1–12, May 2013.
[46] D. Hong, J. Yao, C. Li, D. Meng, N. Yokoya, and J. Chanussot, "Decoupled-and-coupled networks: Self-supervised hyperspectral image super-resolution with subpixel fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5527812.
[47] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, "Extended vision transformer (ExViT) for land use and land cover classification: A multimodal deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5514415.
[48] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Syst.*, vol. 16, no. 6, pp. 345–379, Nov. 2010.
[49] R. Daza, A. Morales, R. Tolosana, L. F. Gomez, J. Fierrez, and J. Ortega-Garcia, "EdBB-demo: Biometrics and behavior analysis for online educational platforms," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 16422–16424.
[50] G. Rebolledo-Mendez, I. Dunwell, E. A. Martínez-Mirón, M. D. Vargas-Cerdán, S. D. Freitas, F. Liarokapis, and A. R. García-Gaona, "Assessing Neurosky's usability to detect attention levels in an assessment exercise," in *Proc. Intl. Conf. Human-Comput. Interact.*, 2009, pp. 149–158.
[51] F.-R. Lin and C.-M. Kao, "Mental effort detection using EEG data in e-learning contexts," *Comput. Educ.*, vol. 122, pp. 63–79, Jul. 2018.
[52] J. Yao, X. Cao, D. Hong, X. Wu, D. Meng, J. Chanussot, and Z. Xu, "Semi-active convolutional neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5537915.
[53] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5202–5211.
[54] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5525–5533.
[55] X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Style aggregated network for facial landmark detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 379–388.
[56] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image Vis. Comput.*, vol. 47, pp. 3–18, Mar. 2016.
[57] J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Target dependent score normalization techniques and their application to signature verification," *IEEE Trans. Syst., Man Cybern., C*, vol. 35, no. 3, pp. 418–425, Aug. 2005.
[58] L. F. Gomez, A. Morales, J. Fierrez, and J. R. Orozco-Arroyave, "Exploring facial expressions and action unit domains for Parkinson detection," *PLoS ONE*, vol. 18, no. 2, Feb. 2023, Art. no. e0281248.
[59] L. F. Gomez, A. Morales, J. R. Orozco-Arroyave, R. Daza, and J. Fierrez, "Improving Parkinson detection using dynamic features from evoked expressions in video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2021, pp. 1562–1570.
[60] R. Berral-Soler, F. J. Madrid-Cuevas, R. Muñoz-Salinas, and M. J. Marín-Jiménez, "RealHePoNet: A robust single-stage ConvNet for head pose estimation in the wild," *Neural Comput. Appl.*, vol. 33, no. 13, pp. 7673–7689, Jul. 2021.
[61] N. Gourier, D. Hall, and J. L. Crowley, "Estimating face orientation from robust detection of salient facial structures," in *Proc. FG Net Workshop Vis. Observ. Deictic Gestures*, 2004, p. 7.
[62] M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 2144–2151.
[63] J. Fierrez-Aguilar, L. Nanni, J. Lopez-Penalba, J. Ortega-Garcia, and D. Maltoni, "An on-line signature verification system based on fusion of local and global information," in *Proc. Intl. Conf. Audio Video-Based Biometric Person Authentication*, 2005, pp. 523–532.
[64] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, and J. Ortega-Garcia, "Feature-based dynamic signature verification under forensic scenarios," in *Proc. 3rd Int. Workshop Biometrics Forensics (IWBF)*, Mar. 2015, pp. 1–6.
[65] W. Zhang, X. Ji, K. Chen, Y. Ding, and C. Fan, "Learning a facial expression embedding disentangled from identity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6755–6764.
[66] R. Vemulapalli and A. Agarwala, "A compact embedding for facial expression similarity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5676–5685.
[67] W. Chen and D. McDuff, "DeepPhys: Video-based physiological measurement using convolutional attention networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 349–365.
[68] J. Hernandez-Ortega, J. Fierrez, A. Morales, and D. Diaz, "A comparative evaluation of heart rate estimation methods using face videos," in *Proc. IEEE 44th Annu. Comput., Softw., Appl. Conf. (COMPSAC)*, Jul. 2020, pp. 1438–1443.
[69] G. Heusch, A. Anjos, and S. Marcel, "A reproducible study on remote heart rate measurement," 2017, *arXiv:1709.00962*.
[70] J. Huertas-Tato, A. Martín, J. Fierrez, and D. Camacho, "Fusing CNNs and statistical indicators to improve image classification," *Inf. Fusion*, vol. 79, pp. 174–187, Mar. 2022.
[71] J. K. Portello, M. Rosenfield, and C. A. Chu, "Blink rate, incomplete blinks and computer vision syndrome," *Optometry Vis. Sci.*, vol. 90, no. 5, pp. 482–487, 2013.
[72] A. Abusharha, "Changes in blink rate and ocular symptoms during different reading tasks," *Clin. Optometry*, vol. 9, pp. 133–138, Nov. 2017.
[73] M. Navarro, A. Becerra, R. Daza, R. Cobos, A. Morales, and J. Fierrez, "VAAD: Visual attention analysis dashboard applied to e-learning," in *Proc. Int. Symp. Comput. Educ. (SIIE)*, 2024, pp. 1–6.
[74] A. Morales, J. Fierrez, R. Tolosana, J. Ortega-Garcia, J. Galbally, M. Gomez-Barrero, A. Anjos, and S. Marcel, "Keystroke biometrics ongoing competition," *IEEE Access*, vol. 4, pp. 7736–7746, 2016.
[75] A. Morales, J. Fierrez, M. Gomez-Barrero, J. Ortega-Garcia, R. Daza, J. V. Monaco, J. Montalvão, J. Canuto, and A. George, "KBOC: Keystroke biometrics OnGoing competition," in *Proc. IEEE 8th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Fier, Monaco, Sep. 2016, pp. 1–6.
[76] G. Stragapede, R. Vera-Rodriguez, R. Tolosana, A. Morales, N. Damer, J. Fierrez, and J. Ortega-Garcia, "Keystroke verification challenge (KVC): Biometric and fairness benchmark evaluation," *IEEE Access*, vol. 12, pp. 1102–1116, 2024.

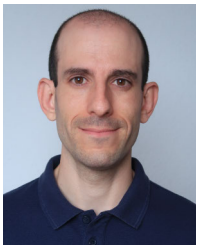
ROBERTO DAZA received the bachelor's degree in telecommunications technology engineering from Universidad de Granada, in 2016, and the M.Sc. degree in telecommunication engineering from Universidad Oberta de Catalunya and the Universitat Ramon Llull-La Salle, in 2019. He is currently pursuing the Ph.D. degree with the Biometrics and Data Pattern Analytics Laboratory, Universidad Autonoma de Madrid. He has participated in several National and European projects focused on the improvements of e-learning, health technologies, and security. His research interests include human–machine interaction, machine learning, deep learning, and biometrics signal processing, with an emphasis on e-learning technologies for security and learning improvement. He has received awards from the eMadrid network, AERFAI, Universidad Autonoma de Madrid, and doctoral schools all over Madrid.

LUIS F. GOMEZ received the bachelor's and M.Sc. degrees in telecommunications engineering from Universidad de Antioquia, Medellin, Colombia, in 2018 and 2021, respectively. He is currently pursuing the Ph.D. degree with the BiDA Lab, Universidad Autonoma de Madrid. During the last five years, he has performed research activities on signal processing image processing, pattern recognition, machine learning, and deep learning, with a focus on biometrics signal processing and their health care and security applications with academic and industrial partners.

JULIAN FIERREZ (Member, IEEE) received the M.Sc. and Ph.D. degrees in telecommunications engineering from Universidad Politecnica de Madrid, Spain, in 2001 and 2006, respectively. Since 2002, he has been with the Universidad Politecnica de Madrid. Since 2004, he has been with Universidad Autonoma de Madrid, where he has been a Full Professor, since 2022. From 2007 to 2009, he was a Visiting Researcher with Michigan State University, USA, under a Marie Curie Fellowship. He was a recipient of a number of world-class research distinctions, including the EBF European Biometric Industry Award 2006, the EURASIP Best Ph.D. Award, in 2012, the Medal in the Young Researcher Award by the Spanish Royal Academy of Engineering, in 2015, and the Miguel Catalan Award. He has been the General Chair of the IAPR Iberoamerican Congress on Pattern Recognition (CIARP 2018) and the Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA 2019). Since 2016, he has been an Associate Editor of *Information Fusion* (Elsevier) and IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY. Since 2018, he has been an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING.



AYTHAMI MORALES received the M.Sc. degree in electrical engineering and the Ph.D. degree in artificial intelligence from Universidad de Las Palmas de Gran Canaria, in 2006 and 2011, respectively. He has performed research stays at the Biometric Research Laboratory, Michigan State University; the Biometric Research Center, The Hong Kong Polytechnic University; the Biometric System Laboratory, University of Bologna; and the Schepens Eye Research Institute (Harvard Medical School). He performs his research works with the Biometric and Data Pattern Analytics Laboratory (BiDA Lab), Universidad Autonoma de Madrid, where he is currently an Associate Professor (CAM Lecturer Excellence Program). He is a member of the European Laboratory for Learning and Intelligent Systems (ELLIS Society). He is the author of more than 100 scientific articles published in international journals and conferences. He holds two patents. His work was supported by the Madrid Government (Comunidad de Madrid-Spain) under the Multiannual Agreement with Universidad Autonoma de Madrid in the line of Excellence for the University Teaching Staff in the context of the Regional Program of Research and Technological Innovation (V PRICIT).



RUBEN TOLOSANA received the M.Sc. degree in telecommunication engineering and the Ph.D. degree in computer and telecommunication engineering from Universidad Autonoma de Madrid, in 2014 and 2019, respectively. In 2014, he joined the Biometrics and Data Pattern Analytics Laboratory (BiDA Lab), Universidad Autonoma de Madrid, where he is currently an Assistant Professor. He is the author of more than 80 scientific papers published in international journals and conferences. His research interests include signal and image processing, pattern recognition, and machine learning, particularly in the areas of DeepFakes, human–computer interaction, biometrics, and health. He is a member of the ELLIS Society, Technical Area Committee of EURASIP, and an Editorial Board of the IEEE Biometrics Council Newsletter. He has also received several awards, such as the European Biometrics Industry Award from the European Association for Biometrics (EAB), in 2018, and the Best Ph.D. Thesis Award from the Spanish Association for Pattern Recognition and Image Analysis (AERFAI), in 2019 and 2022. He has served as the General Chair and the Program Chair for AVSS 2022 and an Area Chair for IJCB 2023 and ICPR 2022 in top conferences.

JAVIER ORTEGA-GARCIA (Fellow, IEEE) received the M.Sc. degree in electrical engineering and the Ph.D. degree (cum laude) in electrical engineering from Universidad Politecnica de Madrid, Spain, in 1989 and 1996, respectively. He is currently a Full Professor with the Signal Processing Chair, Universidad Autonoma de Madrid, Spain, where he holds courses on biometric recognition and digital signal processing. He is also the Founder and the Director of the BiDA Lab, Biometrics and Data Pattern Analytics Group. He has authored over 300 international contributions, including book chapters, refereed journal articles, and conference papers. His research interests include biometric pattern recognition (online signature verification, speaker recognition, and human-device interaction) for security, e-health, and user profiling applications. He chaired Odyssey-04, The Speaker Recognition Workshop, ICB-2013, the 6th IAPR International Conference on Biometrics, ICCST 2017, and the 51st IEEE International Carnahan Conference on Security Technology.