A Machine Learning based Real-Time Application for Engagement Detection

Emanuele Iacobelli[1], Samuele Russo[2] and Christian Napoli[1,3,4]

[1]Department of Computer, Control and Management Engineering, Sapienza University of Rome, 00185 Roma, Italy;
[2]Department of Psychology, Sapienza University of Rome, 00185 Roma, Italy;
[3]Institute for Systems Analysis and Computer Science, Italian National Research Council, 00185 Roma, Italy;
[4]Department of Computational Intelligence, Czestochowa University of Technology, 42-201 Czestochowa, Poland;

Abstract

The study of human engagement has significantly grown in recent years, particularly accelerated by the interaction with a growing number of smart computing machines [1, 2, 3]. Engagement estimation has significant importance across various domains of study, including advertising, marketing, human-computer interaction, and healthcare [4, 5, 6]. In this paper, we propose a real-time application that leverages a single RGB camera to capture user behavior. Our approach implements a novel method for estimating human engagement in real-world scenarios by extracting valuable information from the combination of facial expressions and gaze direction analysis. To acquire this data, we employed fast and accurate machine learning algorithms from the external library dlib, along with custom versions of Residual Neural Networks implemented from scratch. For training our models, we used a modified version of the DAiSEE dataset, a multi-label user affective states classification dataset that collects frontal videos of 112 different people recorded in real-world scenarios. In the absence of a baseline for comparing the results obtained by our application, we conducted experiments to assess its robustness in estimating engagement levels, leading to very encouraging results.

Keywords

Engagement Detection, Eye Tracking, Face Expression Recognition, Machine Learning, Residual Neural Networks

1. Introduction

In today's rapidly evolving digital landscape, humanity interacts with a growing number of smart computing machines. This situation highlights the increasing trend of direct interactions with smart devices in various domains, including household assistance, customer service, and industrial applications. Despite this technological advancement, many devices lack algorithms capable of perceiving and responding to users' attentional states. Traditional user interfaces still heavily rely on explicit input or predefined triggers, resulting in often inefficient and mechanical interactions.

The potential for automatic acquisition and interpretation of users' engagement represents a huge usability improvement for Human-Computer Interaction (HCI) and Human-Robot Interaction (HRI) systems. This capability holds the promise of ushering in more advanced and intuitive interactions, elevating system responsiveness, and enhancing overall user experience. In detail, engagement is a fundamental aspect of the human experience and captures in depth the quality of an individual's involvement, focus, and interaction with their surroundings. For detecting it, facial expressions and gaze direction are crucial elements. In particular, the motion of the eyes is an important element to employ since it highlights the psychological mechanisms behind the human mind and naturally gravitates toward objects, people, or specific regions of interest in the environment.

In this paper, we propose a real-time application that combines gaze direction and face expression analysis to determine the engagement level of a person while interacting with intelligent systems. To achieve this, we defined two machine-learning pipelines leveraging RGB videos of a person interacting with the system. The first pipeline focuses on the user's facial expressions analysis and employs a residual neural network architecture. The second pipeline concentrates on the user's gaze direction estimation by combining pre-trained face and facial landmark detection models with a fast computer vision algorithm that we developed. Predictions of the user's engagement level are ultimately calculated by merging the outputs of these two pipelines using a weighted linear interpolation formula.

Addressing the challenge posed by the absence of a baseline for reference, our primary hurdle in handling this task involved creating an appropriate dataset for training our models. We opted to customize the Affective States in E-Environment Dataset (DAiSEE) [7], a comprehensive collection of multi-label videos designed for identifying user affective states. Given that the estimation of

SYSYEM 2023: 9th Scholar's Yearly Symposium of Technology, Engineering and Mathematics, Rome, December 3-6, 2023
iacobelli@diag.uniroma1.it (E. Iacobelli); samuele.russo@uniroma1.it (S. Russo); cnapoli@diag.uniroma1.it (C. Napoli)
0009-0003-1379-9106 (E. Iacobelli); 0000-0002-1846-9996 (S. Russo); 0000-0002-3336-5853 (C. Napoli)

engagement levels necessitates both temporal and spatial information, videos proved to be an ideal choice. However, to mitigate the high computational and resource costs associated with treating videos as opposed to single images, we implemented mandatory preprocessing steps to optimize memory and computational efficiency.

Continuing to tackle the absence of a reference baseline, we conducted experiments to assess the robustness and effectiveness of our application. This evaluation was carried out using quantitative metrics.

1.1. Roadmap

This paper is organized in the following way: first of all, a summary of the state-of-the-art systems and techniques to recognize the human engagement level is presented (see Section 2). Subsequently, a description of the dataset that we have developed for training our models is illustrated (see Section 3). Following this, a detailed overview of the architectures employed for our application is provided (see Section 4). Then, the results obtained by testing our system considering the quantitative metrics are presented (see Section 5). Finally, we summarize the article's content and outline the possible viable improvements that can be made to our application (see Section 6).

2. Related Works

The field of engagement level detection has seen significant growth, particularly fueled by the global pandemic. With many individuals compelled to participate in remote meetings, analyzing engagement in online sessions has become a pivotal focus, leading to the development of numerous systems. Some studies have explored physiological factors like fatigue [8], brain status and data [9, 10], blood flow and heart rate [11], and galvanic skin conductance [12]. However, due to the recent needs and the remote nature of this task, there has been widespread exploration of inexpensive and unobtrusive technologies. Eye trackers [13, 14] and facial expression recognition models [15, 16] using simple RGB cameras are now among the most promising options.

In a comprehensive review treated in [17], the state-of-the-art engagement detection techniques within the context of online learning are explored. The authors classify existing methods into three primary categories: automatic, semi-automatic, and manual. This classification is based on the methods' dependencies on learners' participation. Furthermore, each category is subdivided based on the type of input data used (e.g., audio, video, text). Among these, video-based methods in the automatic category that leverage facial expressions emerge as the most prevalent. These methods are favored for their ease of implementation and their proven effectiveness in achieving accurate results. The prominence of such techniques underscores the significance of visual cues, particularly facial expressions, in gauging user engagement levels during online interactions.

The work presented in [18] investigates the suitability of three popular models: All-CNN [19], NiN-CNN [20], and VD-CNN [21], along with a customized Convolutional Neural Network (CNN) [22] for detecting engagement level of online learners in educational activities. All the analyzed models leverage facial expressions for scalable and accessible engagement detection. Each of the three base models has its distinct features and the customized CNN combines these advantageous features. For instance, by replacing linear convolutional layers with a multilayer perceptron, increasing depth with small convolutional filters, and replacing some max-pooling layers with convolutional layers with increased stride. All the analyzed models were evaluated on the DAiSEE dataset (extensively explained in Section 3) and the results reveal that the customized CNN outperforms the base models in detecting the engagement level.

In a similar study proposed in [23], the automatic recognition of student engagement from facial expressions is examined using a three-stage pipeline. The initial step involves face registration, detection, and the estimation of key facial landmarks (e.g., eyes, nose, and mouth) by using the approach described in [24]. The second stage employs four binary classifiers to classify the cropped face, distinguishing whether it belongs to one of four engagement levels ($l \in 1, 2, 3, 4$), where 1 signifies no engagement and 4 represents full focus. The authors compared three models for the binary classifier: Support Vector Machines with Gabor features (SVM (Gabor)) [24], Multinomial Logistic Regression with expression outputs from the Computer Expression Recognition Toolbox (MLR(CERT)) [24], and GentleBoost with Box Filter features (Boost(BF)) [25]. This study reveals that SVM (Gabor) yields the best results. The third stage integrates the outputs of all four binary classifiers, utilizing a Multinomial Logistic Regressor model to estimate the final engagement level.

In [26], the authors introduced a regression model for predicting engagement level as a single scalar value from RGB video streams captured by two cameras on the torso and head of an autonomous mobile robot, utilized for tours at The Collection museum in Lincoln, UK. The model incorporates CNN and Long Short-Term Memory (LSTM) [27] networks for video data analysis. Training and evaluation of this regressor network were conducted using a dataset built from the recordings of the autonomous tour guide robot in the public museum. The dataset, manually annotated by three independent people, assigns scalar values in the range [0,1] to represent the user's engagement level. The model demonstrates

Figure 1: Some sample instances present in our customized version of the DAiSEE before converting them in grayscale and applying an histogram equalization. From the left to the right, we have: a) very low engaged, b) low engaged, c) highly engaged, and d) very highly engaged.

optimal engagement level predictions, achieving a Mean Squared Error (MSE) prediction loss of up to 0.126 on the test dataset.

The research conducted in [28] focused on investigating the Deep Facial Spatiotemporal Network (DFSTN). Comprising two integral modules, namely the pretrained SE-ResNet-50 (SENet) utilized for extracting facial spatial features and an LSTM network with Global Attention for generating an attentional hidden state, the DFSTN synergistically captures both facial spatial and temporal information. This combined information is crucial for enhancing engagement prediction performance. The model underwent testing on the DAiSEE dataset, achieving an accuracy of 58.84%, showcasing its capability to outperform numerous existing engagement prediction networks trained on the same dataset.

In [29], the estimation of human attention is based on the direction of the user's face, considering five different directions: central, lateral to the left, lateral to the right, towards up, and towards down. If the user looks in any direction other than the central one, they are assumed to be distracted, with only the central gaze indicating full focus. The authors created a dataset for training, comprising 270 videos of approximately 20 seconds each from 18 different individuals. To enhance data diversity, GAN-based data augmentation techniques were employed to generate new samples, diversifying somatic features in the recorded videos. Transfer Learning [30] was utilized to construct the classifier. Specifically, a pre-trained VGG16 [21] architecture was employed, with three additional dense layers attached at the end for attention estimation.

The approach presented in [31] offers a novel method for estimating driver attention. Departing from conventional methods that primarily focus on a single frontal scene image to analyze driver gaze or head pose, this method introduces a dual-view scene. The additional input data includes the frontal view of the car that the driver is observing. Specifically, the gaze direction is detected and transformed into a probability map of the same size as the road view image, while salient features of temporal and spatial dimensions are extracted from the road view images. These features are then combined and fed into a multi-resolution neural network tasked with driver attention estimation, generating a heat map on the images representing the road. The training dataset for this model is constructed using virtual reality and a driving simulator, incorporating images from the DR(eye)VE dataset [32] that depict the frontal view of the road observed by the driver. Experimental results showcase the feasibility and superiority of the proposed method over existing approaches.

3. Dataset

The baseline dataset utilized for training our networks is a customized version of the Dataset for Affective States in E-Environments (DAiSEE), a large collection of multi-label videos designed for identifying user affective states, including boredom, confusion, engagement, and frustration in real-world scenarios. This dataset comprises 9068 frontal view videos featuring 112 distinct individuals expressing different levels of affective states. Each of these states was manually ranked utilizing the following scale: very low, low, high, and very high.

To create our customized dataset we initially modified the task from which DAiSEE was originally built. We switched from multi-label to multi-class classification, associating only the level of engagement with each video and removing the labels for the other affective states. Example instances present inside our customized version of the DAiSEE are displayed in Fig. 1. Subsequently, we divided the dataset into Training, Validation, and Test sets, with proportions of 60%, 20%, and 20%, respectively. However, the resulting sets were highly unbalanced due to a small portion of videos classified as very low and low engagement. To address this issue, we downsampled the dataset in several ways to achieve a more balanced distribution. First of all, redundancy in subjects was reduced by removing multiple videos of the same individuals. Then, through the use of a normal distribution, we sampled the remaining data instances considering the frequency of labels in the videos with the following formula:

$$n_i = f_i \cdot \frac{n_{tot} - f_i}{n_{tot}} \cdot \lambda \tag{1}$$

Table 1
This table displays the number of sample instances before and after the customization of the Training, Validation, and Test sets derived from the DAiSEE.

Engagement Level	Original Dataset			Customized Dataset		
	Training	Validation	Test	Training	Validation	Test
Very Low	34	23	4	34	23	4
Low	214	160	81	52	37	17
High	2649	912	861	341	110	105
Very High	2585	625	777	344	100	102

where λ is the reduction coefficient (that we have set to 0.25), n_{tot} represents the total number of samples in a given set, and f_i denotes the frequency of label i in that set. Table 1 displays information both before and after the preprocessing procedures on the dataset.

Since DAiSEE contains recordings captured in dynamic environments, each of these videos may present different and various disturbances, such as changing light conditions, visual occlusions, or unconstrained user motion. To improve video quality, we applied manual color and intensity adjustments, focusing on enhancing contrast, brightness, and sharpness for optimal detail resolution. Examples include adjustments to the gamma value, which effectively improves visibility in varying light conditions or exposure levels by normalizing image histograms, making videos more suitable for continuous analysis; Another example is the sharpness adjustments, which enhance fine details and edges, making facial features more prominent.

Despite these modifications, the dataset still demanded excessive memory requirements. Consequently, we opted for further adjustments. Considering that the majority of engagement information is likely derived from human expressions and gaze attention, with a smaller contribution from gestures, we decided to crop from each video only the user's faces. This step also aimed to eliminate potential issues and biases arising from background data. The face cropping was automated using a pre-trained Single Shot Multibox Detector (SSD) model from the Caffe framework [33].

To prevent the generation of unstable videos, we applied a stabilization algorithm (see pseudocode in Fig. 2) that facilitates smooth transitions between subsequently detected faces by stabilizing the position of their bounding boxes. At the start of each video, the size of the first detected face's bounding box is stored. In all the following frames, this dimension is used to resize the bounding box of the subsequently detected faces. Additionally, if the distance between the centers of two consecutive detected faces is smaller than a manually adjusted threshold γ, the center of the newest detected face is replaced with the center of the bounding box of the previously detected face. Finally, each frame is converted to grayscale, and histogram equalization is applied to normalize the color information.

Algorithm 1: Face Stabilization

Data: $F_{original}$: set of frames composing a video.
γ: specific value assigned for the input video.
Result: F_{custom}: set of frames containing only stabilized faces.

$prev_c \leftarrow None$
foreach *frame* $f \in F_{original}$ **do**
  $box_f \leftarrow$ bounding_box_face(f)
  if f *is the first frame of* $F_{original}$ **then**
    $w_{box}, h_{box} \leftarrow$ width(box_f), height(box_f)
  end
  $c_f \leftarrow$ calculate_center(box_f)
  $box_f^{'} \leftarrow (c_{f,x} - \frac{w_{box}}{2}, c_{f,y} - \frac{h_{box}}{2}),$
    $(c_{f,x} + \frac{w_{box}}{2}, c_{f,y} + \frac{h_{box}}{2})$
  $c_f^{'} \leftarrow$ calculate_center($box_f^{'}$)
  if $prev_c \neq None$ **then**
    $d = \|c_f^{'} - prev_c\|$
    if $d < \gamma$ **then**
      $c_f^{'} = prev_c$
      $box_f^{'} \leftarrow (c_{f,x}^{'} - \frac{w_{box}}{2}, c_{f,y}^{'} - \frac{h_{box}}{2}),$
        $(c_{f,x}^{'} + \frac{w_{box}}{2}, c_{f,y}^{'} + \frac{h_{box}}{2})$
    end
  end
  $prev_c \leftarrow c_f^{'}$
  $f^{'} \leftarrow$ crop_face($f, box_f^{'}$)
  $F_{custom} \leftarrow$ append($f^{'}$)
end

Figure 2: Pseudocode of the face stabilization algorithm used to prevent unstable videos while cropping the user's faces from the original video in the DAiSEE.

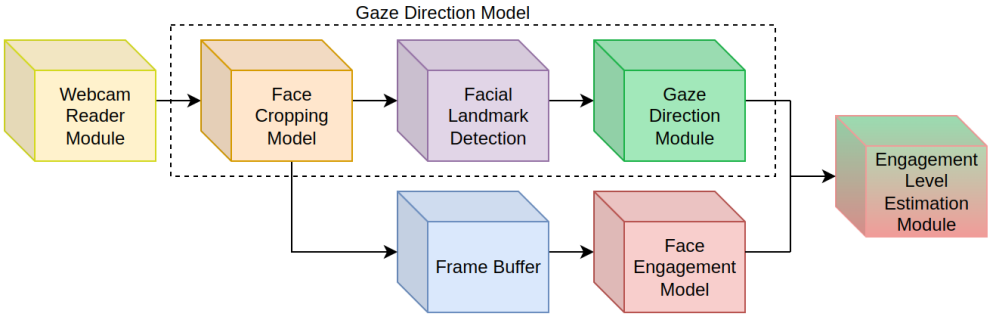

Figure 3: Full pipeline of the real-time application. The Webcam Reader Module acquires the data in real-time and passes them to the Face Cropping Model. This model crops the user's face from the webcam images and passes them both to the Facial Landmark Detection module and the Frame Buffer which has a capacity of 60 frames. Once the buffer is full, each new frame is passed to the Gaze Direction Module and the Face Engagement Model. The predictions of these models are then combined to produce the actual output of our system.

4. Methodology

The complete architecture of the real-time application we developed is illustrated in Fig. 3. Specifically, the system utilizes a single input video stream captured through a webcam reader module, implemented in the external library OpenCV [34], to feed two distinct models. The Face Engagement Model evaluates engagement based on facial expressions, while the Gaze Direction Model predicts engagement by analyzing where the user's focal point. Lastly, the predictions of these two models are combined to derive the final engagement level estimated by our application.

4.1. Face Engagement Model

This model is designed to estimate the user engagement level from frontal recording videos. We designed it as a customized version of the ResNet architecture [35] and we implemented different versions to identify the most effective one. In essence, a residual network employs skip connections to address the vanishing gradient problem. These connections allow information to directly back-propagate, circumventing previous layers. Moreover, a skip connection facilitates a residual block in learning the residual, which is the difference between the desired output and the current input of the layer. This approach makes it easier for the network to understand what input modifications are needed to achieve the desired output, rather than altering the entire input from scratch. This often translates to a more straightforward learning process for the network.

To address the human engagement level classification problem, our models needed to capture both spatial and temporal information. To enable the network to learn temporal information by analyzing multiple frames simultaneously in the same layer, we opted for 3D convolutional layers instead of the traditional 2D convolutional layers implemented in the original ResNet architecture. Learning temporal information is crucial for video analysis, as it allows the network to recognize complex patterns such as actions, gestures, or sequences of facial expressions. Due to this requirement, the model necessitates an initial period to populate a buffer of 60 frames, ensuring a sufficient amount of data for the correct utilization of the 3D convolutional layers. Once the buffer reaches its capacity, the prediction of the engagement level can begin. Subsequently, with the arrival of each new frame, the buffer is updated, and the oldest frame is discarded. We tested three versions of this architecture, differing mainly in the depth and the internal structure of the convolutional block used. Specifically, we implemented the 18-, 34-, and 50-layer versions.

For all these architectures, we introduced 3D layers for batch normalization, max pooling, and average pooling. In detail, each convolutional block includes a batch normalization layer, and all convolutional layers employ the ReLU activation function. Only the last fully connected layer, responsible for the final prediction of the human's engagement level, uses the Softmax activation function. During training, we utilized the He/Kaiming initialization technique [36], which initializes weights using a normal distribution with zero mean and a variance of $\frac{2}{n}$, where n is the total number of inputs to the neuron. This initialization is specifically tailored for networks employing the ReLU activation function, mitigating the vanishing or exploding gradient problem.

Additionally, we employed the Focal Loss [37] as the training function, opting for it over the conventional Categorical Cross-Entropy. The principal reason is that the Focal Loss addresses the issue of unbalanced data by prioritizing examples the model struggles with, rather than those it confidently predicts. This ensures continuous

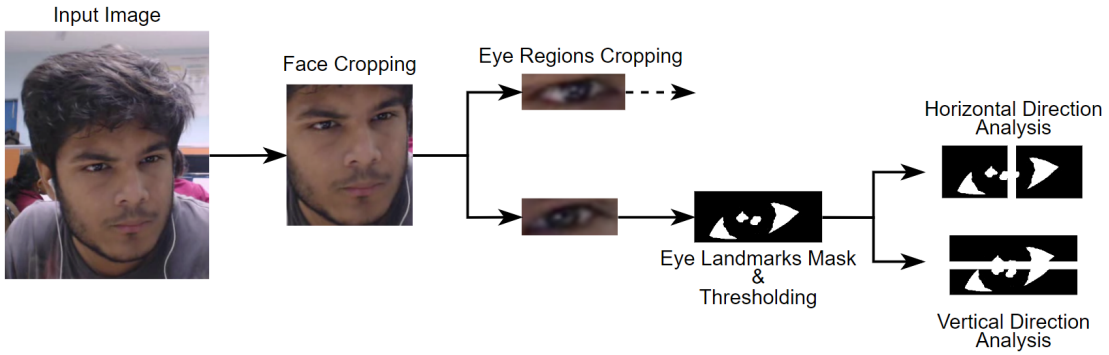

Figure 4: Gaze Direction Model Workflow: The input image, captured by the webcam, undergoes processing through the Face Cropping Module. This module is responsible for cropping the detected face, and the resulting image is then fed into the Face Landmark Detector. The Face Landmark Detector estimates the position of facial keypoints, which are subsequently utilized to crop the eye regions. Each eye image undergoes further analysis in the Gaze Direction Module, which assesses both horizontal and vertical directions of the gaze.

improvement on challenging examples, preventing the model from becoming overly confident with easy ones. We implemented the following Focal Loss formula:

$$-(1-p_i)^{\gamma}\ln(p_i) \tag{2}$$

where γ represents the focusing parameter (typically a positive number) to be fine-tuned using cross-validation, and p_i denotes the predicted probability of the correct class. Also, our training process incorporates early stopping with a learned patience value of 10 epochs, L2 regularization featuring a weight decay set to $1e^{-3}$, and an Adam Optimizer accompanied by a Learning Rate Scheduler [38] with a maximum learning rate of $1e-4$ and a Gradient Scaler to reduce the range of magnitudes in the gradients. All the implementation details of the tested models are reported in the Table 2. Following the training phase, we opted for the 50-layer version model, with a batch size equal to 16, as the engagement network for our application, as it demonstrated the highest accuracy among the tested versions.

4.2. Gaze Direction Model

This model is designed to extract attention information from a person's gaze in frontal recording videos. The gaze direction provides valuable insights into a person's engagement during a task. The complete workflow of this model is displayed in Fig. 4. To implement this model, we combined two pre-trained neural networks available in the dlib library [39].

The first network is the Face Cropping Model, a CNN trained for face detection in general images. It not only identifies faces but also provides their bounding box coordinates and converts the input image to grayscale. Although the use of this network may appear redundant considering the customized dataset that we have employed for training the engagement model, it plays a crucial role in the real-time application. Specifically, it crops faces from the live stream frames and passes these images to both the Engagement Model and the Facial Landmark Detector.

The Facial Landmark Detector, the second network

Table 2
This table displays the implementation details of the three different architectures that we tested for the Face Engagement Model. When the stride information is not present it means that the stride is equal to 1.

Layer Name	Architecture Name		
	18-layer	34-layer	50-layer
Convolution	kernel = [7,7,7], filters = 64, stride = 2		
Max Pool	k = [3,3,3], s = 2		
Convolution Block	k=[3,3,3],f=64 k=[3,3,3],f=64 x2	k=[3,3,3],f=64 k=[3,3,3],f=64 x3	k=[1,1,1],f=64 k=[3,3,3],f=64 k=[1,1,1],f=256 x3
Convolution Block	k=[3,3,3],f=128 k=[3,3,3],f=128 x2	k=[3,3,3],f=128 k=[3,3,3],f=128 x4	k=[1,1,1],f=128 k=[3,3,3],f=128 k=[1,1,1],f=512 x4
Convolution Block	k=[3,3,3],f=256 k=[3,3,3],f=256 x2	k=[3,3,3],f=256 k=[3,3,3],f=256 x6	k=[1,1,1],f=256 k=[3,3,3],f=256 k=[1,1,1],f=1024 x6
Convolution Block	k=[3,3,3],f=512 k=[3,3,3],f=512 x2	k=[3,3,3],f=512 k=[3,3,3],f=512 x3	k=[1,1,1],f=512 k=[3,3,3],f=512 k=[1,1,1],f=2048 x3
Average Pool	Output Size = 1x1x1		
Dropout	Rate = 0.4		
Linear	Neurons = 1024		

that we have employed from the dlib library, recognizes 68 2D facial landmarks (e.g., nose tip, corners of the mouth, and eyes) in a given face image. These facial landmarks serve two purposes: they are used to crop the eye regions based on the eye landmarks and to calculate the face orientation with respect to the vertical axis (yaw angle). This orientation is determined through the use of a vector starting from the midpoint between the eyes and terminating at the nose tip.

Estimating the focal point of the user is accomplished through the Gaze Direction Module, a simple computer vision pipeline. Initially, the eye landmarks outlining the eye contours are employed to create a mask that removes extraneous pixels from each cropped eye image. Subsequently, the Otsu's method [40] is applied to automatically threshold the image, distinguishing between foreground (iris and pupil pixels) and background (sclera pixels).

The resulting image is then horizontally and vertically divided around its center to estimate the gaze direction. Both vertical and horizontal gaze directions are quantified as values within the range of [-1,1]. Regarding horizontal gaze direction, a value approaching -1 indicates the user is looking to the left, while a value approaching 1 suggests a rightward gaze. A value around 0 indicates the user is looking at the center of the screen. Similarly, for vertical gaze direction, a value nearing 1 signifies a downward gaze and a value nearing -1 indicates an upward gaze.

To compute these directions, the density of white pixels representing the sclera is analyzed. For each eye image, the total number of white pixels is calculated. If this value is zero, it implies incorrect eye detection, and the current frame is skipped. Otherwise, for each sub-image generated, the percentage of white pixels in relation to the total number of white pixels in the corresponding original eye image is calculated. Then, the percentages belonging to the same direction of both eyes are averaged (e.g., the percentage of white pixels in the left sub-image of the left eye is averaged with the percentage of white pixels in the left sub-image of the right eye). Finally, the difference between these averages produces the value within the range of [-1,1] described earlier.

To effectively use the estimated gaze direction, it's crucial to consider the limits of the user's field of view, which may vary based on the task. In our screen-based task implementation, we assume that a face orientation deviation exceeding 30 degrees from the camera-aligned orientation indicates the user is no longer looking at the monitor.

Initially, these limits are set at the task's beginning and dynamically adjusted based on the user's face position and orientation relative to the camera frame's center. If the face orientation exceeds 20 degrees from the frontal position, the horizontal limits shift proportionally based on the sine of the face orientation. Updates related to face position involve calculating the distance between the face bounding box center and the frame center. If this distance exceeds one-sixth of the total frame dimension, the right and left limits are adjusted. The adjustment is determined by normalizing the distance between the face and frame centers between 0 and 0.5. If the face shifts to the right, the distance is subtracted from the limits; otherwise, it is added.

Table 3
Table displaying the conversion rules from engagement level labels to score and vice versa.

Engagement Label	*Engagement Score*
0	0.1
1	0.35
2	0.65
3	0.9

The engagement level, derived from the gaze direction, is within the range [0,1]. It is obtained by subtracting the sum of horizontal and vertical gaze errors from 1. A score of 1 indicates complete focus on the screen, with no gaze exceeding the defined limits. A score of 0 implies no face detection in the current frame. The closer the engagement level is to zero, the more the user surpasses the admissible field of view limits, indicating a lack of focus on the task. Specifically, when the gaze exceeds the limits, horizontal and vertical gaze errors are calculated as the difference in modulo between the estimated gaze direction and the corresponding limits.

4.3. Engagement Level Estimation

To obtain the final detected engagement level, we combined the predictions from the Face Engagement Model and the Gaze Direction Model using a linear interpolation formula:

$$\alpha \cdot Engagement_{Face} + (1 - \alpha) \cdot Engagement_{Gaze} \quad (3)$$

Where α is a learnable parameter used to weigh the importance of the models' predictions. In addition, to correctly apply this formula, the prediction of the Face Engagement Model needs to be converted from labels to a value within the range [0,1]. The conversion is performed according to the rules displayed in Table 3.

5. Results

To evaluate the accuracy of our system, we measured the disparity between the predicted engagement scores and the ground truth values using two regression metrics: Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). To facilitate the application of these

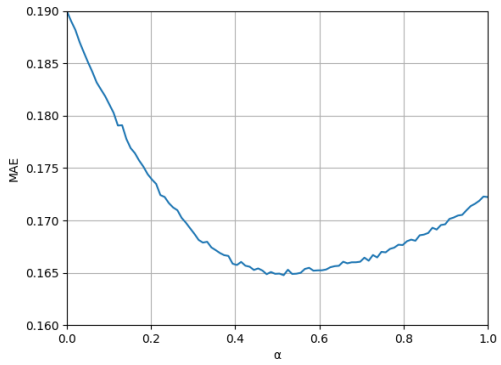

Figure 5: Trend of the Mean Absolute Error (MAE) with varying values of the parameter α in Eq. (3).

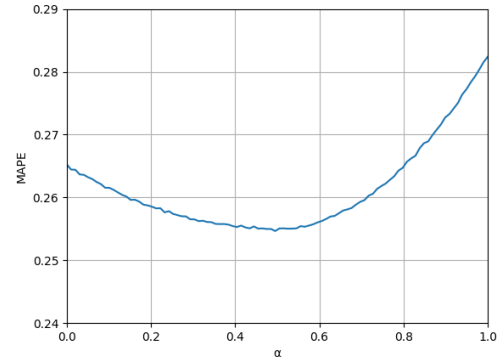

Figure 6: Trend of the Mean Absolute Percentage Error (MAPE) with varying values of the parameter α in Eq. (3).

metrics, we converted the engagement level labels associated with the samples in our customized dataset using the conversion rules outlined in Table 3. This transformation effectively turned the multi-label class problem, designed for the DAiSEE dataset, into a regression problem.

During training, we experimented with different values for the parameter α in Eq. (3) to maximize the system's accuracy. As illustrated in Figs. 5 and 6, the lowest error for both MAE and MAPE occurred when α was set to 0.5. This indicates that both predictions from the Face Engagement Model and the Gaze Direction Model carry equal importance and are essential for achieving accurate predictions.

Analysis of the scenarios where α is 0 (using only the Face Model) or 1 (using only the Gaze Model) reveals significantly higher errors in both performance metrics. Independently, these predictions struggle to accurately gauge the user's engagement level. With α initialized to 0.5, our system achieved an accuracy of approximately 58% (57.7%), closely aligning with the performance of state-of-the-art works in engagement level detection discussed in Section 2 that work with the original version of the DAiSEE.

6. Conclusions

Our work introduces a novel approach to engagement level estimation by integrating two distinct machine learning pipelines focused on analyzing facial expressions and gaze direction. Noteworthy is our real-time application's emphasis on cost-effectiveness and accessibility, achieved through the utilization of a single RGB camera, fast and lightweight machine learning algorithms, and computationally efficient computer vision techniques.

In terms of system training, we customized the DAiSEE dataset to optimize memory usage, reduce class imbalance, mitigate bias introduced by repeated instances of the same individuals, and focus exclusively on facial cropping to eliminate potential background-related biases. The achieved results underscore the potential of our system as a robust foundation, offering a secure benchmark for the development of innovative applications integrating automatic user engagement recognition, thereby dynamically adapting to user interactions. This not only enhances overall usability but also heralds a new era in application interfaces, promising heightened levels of user experience and interaction.

Looking forward, future improvements to our system can be directed towards enhancing the accuracy, robustness, and generalization capabilities by expanding the dataset's dimensions. This expansion may involve incorporating data from a more diverse group, encompassing individuals with varying demographic characteristics, cultural backgrounds, and engagement patterns.

Also, exploring attention estimation in multi-face contexts, where multiple individuals are present simultaneously, represents another intriguing avenue for future research. Lastly, a significant refinement to our application involves substituting the CNN layers in the Face Detection Model with Visual Transformers [41](ViTs), known for their excellence in image manipulation and long-range dependency modeling compared to traditional convolutional layers. This substitution could enhance the precision of engagement level estimation from facial expressions, as different facial regions can be effectively combined at the same time.

References

[1] G. Capizzi, G. L. Sciuto, C. Napoli, M. Woźniak, G. Susi, A spiking neural network-based long-

term prediction system for biogas production, Neural Networks 129 (2020) 271 – 279. doi:10.1016/j.neunet.2020.06.001.

[2] N. Brandizzi, S. Russo, G. Galati, C. Napoli, Addressing vehicle sharing through behavioral analysis: A solution to user clustering using recency-frequency-monetary and vehicle relocation based on neighborhood splits, Information (Switzerland) 13 (2022). doi:10.3390/info13110511.

[3] A. Alfarano, G. De Magistris, L. Mongelli, S. Russo, J. Starczewski, C. Napoli, A novel convmixer transformer based architecture for violent behavior detection 14126 LNAI (2023) 3 – 16. doi:10.1007/978-3-031-42508-0_1.

[4] G. Capizzi, G. L. Sciuto, C. Napoli, E. Tramontana, A multithread nested neural network architecture to model surface plasmon polaritons propagation, Micromachines 7 (2016). doi:10.3390/mi7070110.

[5] C. Napoli, G. Pappalardo, E. Tramontana, R. K. Nowicki, J. T. Starczewski, M. Woźniak, Toward work groups classification based on probabilistic neural network approach, volume 9119, 2015, pp. 79 – 89. doi:10.1007/978-3-319-19324-3_8.

[6] N. Brandizzi, S. Russo, R. Brociek, A. Wajda, First studies to apply the theory of mind theory to green and smart mobility by using gaussian area clustering, volume 3118, 2021, pp. 71 – 76.

[7] A. Gupta, A. D'Cunha, K. Awasthi, V. Balasubramanian, Daisee: Towards user engagement recognition in the wild, arXiv preprint arXiv:1609.01885 (2016).

[8] Z. Wan, J. He, A. Voisine, An attention level monitoring and alarming system for the driver fatigue in the pervasive environment, in: Brain and Health Informatics: International Conference, BHI 2013, Maebashi, Japan, October 29-31, 2013. Proceedings, Springer, 2013, pp. 287–296.

[9] V. Ponzi, S. Russo, A. Wajda, R. Brociek, C. Napoli, Analysis pre and post covid-19 pandemic rorschach test data of using em algorithms and gmm models, volume 3360, 2022, pp. 55 – 63.

[10] C.-M. Chen, J.-Y. Wang, C.-M. Yu, Assessing the attention levels of students by using a novel attention aware system based on brainwave signals, British Journal of Educational Technology 48 (2017) 348–369.

[11] S. Di Palma, A. Tonacci, A. Narzisi, C. Domenici, G. Pioggia, F. Muratori, L. Billeci, M. S. Group, et al., Monitoring of autonomic response to sociocognitive tasks during treatment in children with autism spectrum disorders by wearable technologies: A feasibility study, Computers in biology and medicine 85 (2017) 143–152.

[12] O. Dehzangi, C. Williams, Towards multi-modal wearable driver monitoring: Impact of road condition on driver distraction, in: 2015 IEEE 12th international conference on wearable and implantable body sensor networks (BSN), IEEE, 2015, pp. 1–6.

[13] E. Iacobelli, V. Ponzi, S. Russo, C. Napoli, Eye-tracking system with low-end hardware: Development and evaluation, Information 14 (2023) 644.

[14] F. Fiani, S. Russo, C. Napoli, An advanced solution based on machine learning for remote emdr therapy, Technologies 11 (2023). doi:10.3390/technologies11060172.

[15] P. Kaur, K. Krishan, S. K. Sharma, T. Kanchan, Facial-recognition algorithms: A literature review, Medicine, Science and the Law 60 (2020) 131–139.

[16] G. De Magistris, M. Romano, J. Starczewski, C. Napoli, A novel dwt-based encoder for human pose estimation, volume 3360, 2022, pp. 33 – 40.

[17] M. Dewan, M. Murshed, F. Lin, Engagement detection in online learning: a review, Smart Learning Environments 6 (2019) 1–20.

[18] M. Murshed, M. A. A. Dewan, F. Lin, D. Wen, Engagement detection in e-learning environments using convolutional neural networks, in: 2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), IEEE, 2019, pp. 80–86.

[19] J. T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net, arXiv preprint arXiv:1412.6806 (2014).

[20] M. Lin, Q. Chen, S. Yan, Network in network, arXiv preprint arXiv:1312.4400 (2013).

[21] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[22] S. Albawi, T. A. Mohammed, S. Al-Zawi, Understanding of a convolutional neural network, in: 2017 international conference on engineering and technology (ICET), Ieee, 2017, pp. 1–6.

[23] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, J. R. Movellan, The faces of engagement: Automatic recognition of student engagementfrom facial expressions, IEEE Transactions on Affective Computing 5 (2014) 86–98.

[24] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, M. Bartlett, Computer expression recognition toolbox, Proc. Automatic Face and Gesture Recognition (FG'11) 20 (2011) 24–25.

[25] P. Viola, M. Jones, et al., Robust real-time object detection, International journal of computer vision 4 (2001) 4.

[26] F. Del Duchetto, P. Baxter, M. Hanheide, Are you still with me? continuous engagement assessment from a robot's point of view, Frontiers in Robotics

and AI 7 (2020) 116.
[27] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.
[28] J. Liao, Y. Liang, J. Pan, Deep facial spatiotemporal network for engagement prediction in online learning, Applied Intelligence 51 (2021) 6609–6621.
[29] S. Pepe, S. Tedeschi, N. Brandizzi, S. Russo, L. Iocchi, C. Napoli, Human attention assessment using a machine learning approach with gan-based data augmentation technique trained using a custom dataset, OBM Neurobiology 6 (2022) 1–10.
[30] S. J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on knowledge and data engineering 22 (2009) 1345–1359.
[31] Z. Hu, C. Lv, P. Hang, C. Huang, Y. Xing, Data-driven estimation of driver attention using calibration-free eye gaze and scene features, IEEE Transactions on Industrial Electronics 69 (2021) 1800–1808.
[32] A. Palazzi, D. Abati, F. Solera, R. Cucchiara, et al., Predicting the driver's focus of attention: the dr (eye) ve project, IEEE transactions on pattern analysis and machine intelligence 41 (2018) 1720–1733.
[33] E. Cengil, A. Çınar, E. Özbay, Image classification with caffe deep learning framework, in: 2017 International Conference on Computer Science and Engineering (UBMK), IEEE, 2017, pp. 440–444.
[34] I. Culjak, D. Abram, T. Pribanic, H. Dzapo, M. Cifrek, A brief introduction to opencv, in: 2012 proceedings of the 35th international convention MIPRO, IEEE, 2012, pp. 1725–1730.
[35] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[36] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1026–1034.
[37] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.
[38] L. N. Smith, N. Topin, Super-convergence: Very fast training of neural networks using large learning rates, in: Artificial intelligence and machine learning for multi-domain operations applications, volume 11006, SPIE, 2019, pp. 369–386.
[39] D. E. King, Dlib-ml: A machine learning toolkit, The Journal of Machine Learning Research 10 (2009) 1755–1758.
[40] N. Otsu, A threshold selection method from gray-level histograms, IEEE transactions on systems, man, and cybernetics 9 (1979) 62–66.
[41] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).