Interlcutor's attention-engagement estimation for HRI

Fabrizio Casadei

Abstract

The estimation of human attention is a crucial aspect in various fields, such as human-computer interaction, advertising, and healthcare. In this paper, we propose a novel approach for attention estimation in real-world scenarios using computer vision techniques. By combining facial and gaze analysis, we develop a robust system that accurately assesses attention levels. This paper presents a novel machine learning model for acquiring the level of engagement and the estimation of interlocutor's attention via gaze-related inference. The contributions of this paper are quadruple: (1) We propose variations of *DAiSEE* dataset collected in real-world scenarios to train and evaluate our model. (2) A novel architecture is presented, that integrates gaze-related inference and engagement estimation. (3) We conduct experiments and demonstrated the effectiveness and robustness of our model in acquiring attention and estimating engagement levels. (4) A real-time application that utilizes the methodology outlined has been introduced.

1. Highlights

- Novel approach for attention estimation combining gaze-related attention and engagement (psychological).
- New Custom *ResNet* version for spatiotemporal dynamics, including an experimental learning phase.
- Innovative data processing to improve operations on video sources, balancing, enhancing, and optimizing data storage.
- Introduced a combination of state-of-the-art models and computer vision techniques to infer gaze direction.
- A new baseline for attention estimation using a mixed solution.

2. Introduction

The last two decades have shown how the use of AI-related systems can help and improve our daily life, automatizing operations and assisting people. This has been possible thanks to the continuous technological progress and research, that collaborates to acquire new capabilities.

Human-Robot Interaction (HRI) and Human-Computer Interaction (HCI) are two of the AI application domains where technology is developing most quickly.

Both domains have a massive usage of computer vision and visual perception techniques to improve the interaction with the user, in order to make the robots or devices smarter and more skilled, to upgrade the interaction with humans.

In today's fast-paced digital world, individuals interact with various computing devices, ranging from smartphones to smart home assistants. Is becoming day by day even more common to have direct interactions with smart robots, from household assistants to customer service and industrial applications.

However, many of these interactions still lack the ability to perceive and respond to users' attentional states. Traditional user interfaces often rely on explicit user input or predefined triggers to initiate actions, which can be inefficient and too mechanical.

The ability to automatically acquire and interpret users' attention has the potential to revolutionize HCI & HRI systems, enabling more advanced and intuitive interaction, and enhancing the system's responsiveness together with overall user experience.

In particular, gaze-related inference and the estimation of engagement have emerged as promising avenues to enhance the interaction between humans and machines. The human gaze plays a vital role in human communication and attention. People naturally direct their gaze in the direction of objects, people, or specific regions of interest in the environment. Thus, integrating gaze-related inference into interactive systems can provide valuable insights into users' attentional focus and intentions.

Another important aspect, that can describe the degree of attention, is the concept of *Engagement*.

The Engagement is a fundamental aspect of human experience, capturing the depth and quality of an individual's involvement, focus, and interaction with their surroundings. It has been widely studied in the last years, nevertheless, its multidimensional nature and underlying psychological mechanisms continue to intrigue researchers. It's a complex and dynamic phenomenon that encompasses an individual's cognitive, emotional, and behavioral responses to stimuli within a specific context.

Engagement plays an important role in HRI & HCI, but

CEUR Workshop Proceedings (CEUR-WS.org)

also it's applied in various other domains, including education, entertainment, and marketing.

2.1. Objective

To the best of our knowledge, no existing work simultaneously addresses the challenges of acquiring the interlocutor's attention via gaze-related inference and estimating the level of engagement using a unified machine-learning framework. In this paper, we propose a novel work that combines gaze-related inference and engagement estimation to determine the attention level of an interlocutor over intelligent systems to effectively interact with the user.
In this way, we expect that our model can provide a more comprehensive insight into the user's cognitive state. To accomplish this task we define two different pipelines.
The first pipeline focuses on the estimation of engagement levels and employs a deep neural network architecture developed from scratch.
In parallel, the second pipeline combines in a novel way, the power of existing models and tools that are computer vision related, in order to evaluate the attention level. To obtain the final outcome, the results from the two pipelines are merged together.
One of the main difficulties for this task is the lack of ad hoc data collections that can be used to estimate engagement. Moreover, considering all the possible aspects that can involve the effective attention level, makes this task really complex to be solved, so it's necessary to relax the problem. Engagement estimation requires temporal information besides the spatial of a single image, which means high quantities of data from a video source. Train a model from scratch with all these data is very expensive, both from a time and computational point of view, thus data processing is required. Also, we have to consider not having similar baselines for the mixed solution, that can be used as a reference for the evaluation. After the description of the data used and the system, we propose a set of quantitative and qualitative results to highlight the work and also present a real-time system that makes use of the architecture.

3. Related work

Several studies have explored the estimation of attention levels during interaction with humans. Duchetto, et al. (2020) [1], address the challenge of capturing the multi-faceted nature of engagement by developing a computational model that computes a single scalar engagement value from standard video streams captured from the robot's perspective. The proposed model utilizes Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to analyze the video data. The model is trained and evaluated using a long-term dataset collected from an autonomous tour guide robot deployed in a public museum. The results demonstrate that the model effectively predicts engagement in the specific application domain of the tour guide robot.
From a group of Sapienza students Sveva Pepe & Simone Tedeschini (2022) [2], we have a system that tries to solve the problem making a simplification on the task. The attention is estimated from the orientation of the face which can be: central, lateral to the left, lateral to the Right, towards Up, towards Down. Considering a high level of interlocutor's attention whether the person has the face oriented centrally. The process consists in doing first a face detection step and then using CNN for the classification. Moreover, they have used a GAN model to increase the number of instances to enrich their dataset built from scratch.
Hu, et al., (2022) [3] presents a method for driver attention estimation in intelligent vehicles. Unlike existing methods that focus solely on scene images or driver's gaze or head pose, the proposed method utilizes a dual-view scene with calibration-free gaze direction. Drawing inspiration from human visual mechanisms, the method extracts low-level features, static visual saliency maps, and dynamic optical flow information as input feature maps. These maps are combined with high-level semantic descriptions and a gaze probability map derived from the gaze direction. Experimental results indicate its feasibility and superiority over existing approaches, and the study provides insights into the influence of environmental factors on the method's performance.

4. Data processing

In this section, are introduced the data involved in the model training for the engagement level estimation and for testing the application using both engagement and gaze analysis data. After the description of the dataset, the full sequence of pre-processing operations is outlined.
The initial data are retrieved from the *DAiSEE* (Dataset for Affective states in E-Environment) [4], a pretty large collection of multi-label video designed for the classification of the user affective states, including boredom, confusion, engagement, and frustration, in real-world scenarios.
DAiSEE provides a dataset comprising 9068 video snippets captured from 112 users. These snippets represent users' affective states and are annotated with four inten-

sity levels of labels: very low, low, high, and very high. The annotations were performed by a crowd and correlated with a gold standard annotation created by expert psychologists.
This data resource enables researchers to explore challenges in feature extraction, context-based inference, and the development of machine learning methods for related tasks. It serves as a valuable resource for advancing research in user engagement recognition and its applications in various domains, such as advertising, healthcare, autonomous vehicles, and e-learning.
The employment of images or videos "in the wild" presents numerous challenges due to the uncontrolled and unpredictable nature of the environment. The difficulties arise from the presence of noise caused by various factors including light conditions, variable camera perspectives, no significant background data, visual occlusions, and data variability.
Also for these reasons, several transformation steps are performed, producing several custom versions of the original data. The video clips are first divided into the training set, validation set, and test set, with the first at approximately 60% and the other two at 20% each of the whole data.
The first operation performed modifies slightly the data collection, converting the associated task from a multi-label classification to a multi-class classification. This is possible considering only the 4 labels relative to the engagement degree.
Image pre-processing plays a crucial role in enhancing the quality of data and improving the performance of machine learning models. The first transformation step takes the videos, represented as sequences of color frames in 480p resolution, applying gamma, saturation, and sharpness adjustment. By adjusting the gamma value, we can effectively enhance the visibility of certain features and details in the image. This is particularly useful when dealing with images captured in challenging lighting conditions or with varying levels of exposure.
Gamma adjustment helps to normalize the image intensities, making them more suitable for subsequent analysis and improving the overall interpretability of the visual cues. The correction of the saturation can enhance or reduce the vividness of colors in the image.
This operation can be beneficial in scenarios where the original dataset exhibits variations in the color distribution or lacks sufficient color contrast.
By applying sharpness adjustment instead, we can enhance the fine details and edges in the image, making them more prominent and well-defined. The application of these techniques has been carefully balanced, since excessive adjustments may introduce artifacts or distortions that could adversely impact the model's performance.
The total dimension of enhanced videos is around 16GB.
The second processing step takes into account the fact of having an highly unbalanced dataset, indeed the samples' labels are distributed like expressed in the table 1.

Table 1
labels distribution in the 1^{nd} version of the new dataset (v1)

	Training	*Validation*	*Test*
`Very low`	34	23	4
Low	214	160	81
High	2649	912	861
Very High	2585	625	777

Moreover, we anticipate that training the model with the complete dataset as it is initially, turned out to be unfeasible, both for the computational resources requested and the time required. Therefore, to balance the data the strategy of downsampling the collection has been preferred.
In addition to the problem exposed, the original present an elevated level of candidate repetition in the consequent clips, producing high redundancy that can be cut off to speed up the training. The data are extracted according to the labels' frequency, using this formula $n_i = f_i \cdot \frac{n_{tot} - f_i}{n_{tot}} \cdot \lambda$, with λ the reduction coefficient set to 0.25, n_{tot} the total number of samples of a certain set, f_i the frequency of the label i in that set.
Then the data instances are sampled using Uniform probability distribution ensuring the obtaining of the new desired number of samples.
After the application of this method to balance and reduce the amount of data, the data collection is composed as shown in the table 2.

Table 2
labels distribution in the 2^{nd} version of the new dataset (v2)

	Training	*Validation*	*Test*
`Very low`	34	23	4
`Low`	51	36	19
`High`	342	107	107
`Very High`	341	99	106

How it's possible to appreciate from these tables, the data unbalancing has been considerably reduced, and ulterior downsampling would reduce too much the amount of data instances
The size of this new data collection is around 2.2 GB.
The next process aims to continue the data reduction, removing useless information or secondary data that slows down the computation. In our engagement level estimation approach, we make the assumption that the face plays a key role in conveying the necessary information for accurate engagement assessment. This assumption is based on the understanding that facial expressions, eye gaze, and other facial cues are strong indicators of

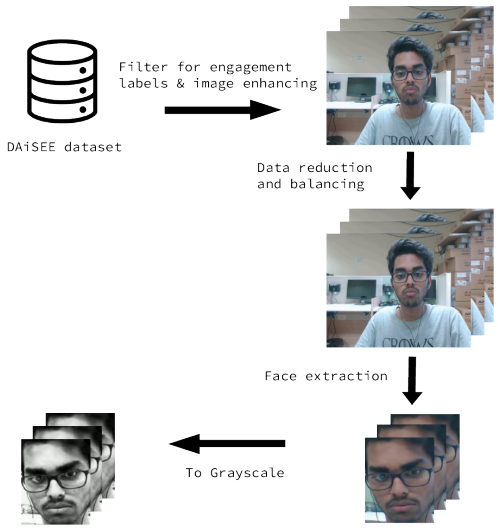


Figure 1: Data elaboration flow, from original *DAiSEE* dataset to the fourth custom version

an individual's level of attention and interest. While the body posture and gestures for example may contain supplementary information, considering them to be of secondary importance in comparison to the rich set of cues available from the face. Besides this, we want also to remove another drawback of this kind of data, the presence of no significant background data, in order to have a better focus.

To perform face detection is exploited a pre-trained deep neural network model from the *Caffe* framework [5]. This pre-trained model is based on the Single Shot Multibox Detector (SSD) framework and is trained to detect faces in images.

To avoid the generation of flickering and unstable video clips a simple algorithm (1) has been applied.

In this algorithm, the γ coefficient regulates when changing the face-framing, setting this to too low results in not resolving the thrill effect. Instead, an excessive value will change the face frame roughly, and also include the possibility to miss part of the face in the bounding box. A coherent value should be chosen accordingly to the context of the videos, i.e. small motions, resolution, subjects' distance, etc. After this process, the dataset size is reduced to 210 MB approximately.

A further version of the data is built upon these last operations, converting the frames from color to grayscale and applying the histogram equalization. The scheme in figure 1 sums up the work performed.

Algorithm 1: Pseudocode for regulation of face extraction in a video clip

Data: Set of Full frames F_s, update threshold γ
Result: Set of Frames with only faces $F_s^{'}$

```
// initialize no previous box center
prev_c ← None
foreach frame F_i ∈ F_s do
    // box from face detection
    box_i ← face detection(F_i)
    // save size of the first frame
    if F_i is first frame then
        w_box, h_box ← width(F_0), height(F_0)
    end
    // center of current box
    c_i ← center(box_i)
    // the new bounding box
    box_i' ← (c_{i,x} − w_box/2, c_{i,y} − h_box/2),
     (c_{i,x} + w_box/2, c_{i,y} + h_box/2)
    // center of new current box
    c_i' ← center(box_i')
    if prev_c ≠ None then
        // difference vector
        v_diff = c_i' - prev_c
        d = ||v_diff||
        if d < γ then
            c_i' = prev_c
            box_i' ← (c_{i,x} − w_box/2, c_{i,y} − h_box/2),
             (c_{i,x} + w_box/2, c_{i,y} + h_box/2)
        end
    end
    // new center as the previous one
    prev_c ← c_i'
    // extract the face frame
    F_i' ← extract(F_i, box_i')
    F_{s,i}' ← F_i'
end
```

5. Methodology

In this section, we point out the whole structure of the application, describing the implementation details and the various components with their role. As introduced in precedence, the system uses two different workflows that are used to provide the estimation of the attention level.

After the discussion on the pipelines, we expose a final treatment on how the relative outcomes are combined.

5.1. Engagement Pipeline

This pipeline presents a custom version of the classical *ResNet* [6] implemented from scratch with different depth levels. To aim of this model is to extract from video sequences the estimation of the user's engagement, during a frontal recording session.

5.1.1. The Model

In our research, we recognize the importance of considering both spatial and temporal information in the data. Training a model on temporal information in addition to spatial information is crucial in various domains, as it enables the model to capture and understand dynamic patterns and changes over time.
While convolutional neural networks with 2D convolutions are effective in extracting spatial features from images, they often overlook the temporal dimension, which is vital for tasks involving video analysis, action recognition, and motion understanding.
By incorporating 3D convolutions into the model architecture, we can effectively process spatiotemporal data. Conv3D operates on a sequence of frames, capturing both the information simultaneously. This allows the model to learn how objects or scenes evolve over time and recognize complex patterns, such as actions, gestures, or object trajectories.
Furthermore, conv3D models excel at handling video sequences with variations in appearance, such as changes in lighting, viewpoint, or background. By leveraging the temporal context, these models can generalize better and exhibit robustness to such variations.
Consequently on this choice, 3D layers for batch normalization, max pooling, and average pooling have been included in the architecture. To reduce the overfitting during training as been included a Dropout layer with a dropout rate of 0.3. In the table 3, are shown the structures for the different types of models that can be utilized in this module.

In Each conv layer is included the batch normalization and the *Relu* as activation function. In the last fully connected layer instead, the activation function for the multi-class problem is the *Softmax*.
This network introduces skip connections, also known as identity mappings, that allow the network to bypass one or more layers and directly propagate information from earlier layers to later layers.
These skip connections enable the network to learn residual mappings, which is the difference between the desired output and the current output of a layer. By learning these residuals, the network can focus on refining the learned representations rather than starting from scratch at each layer.

Table 3
3D Custom Resnet structure

Layer name	*18-layer*	*34-layer*	*50-layer*
`Conv1`	$[7, 7, 7]$, 64, stride 2		
`max pool`	$[3, 3, 3]$, stride 2		
Block Conv1	$[3, 3, 3]$, 64 $[3, 3, 3]$, 64 $\times 2$	$[3, 3, 3]$, 64 $[3, 3, 3]$, 64 $\times 3$	$[1, 1, 1]$, 64 $[3, 3, 3]$, 64 $[1, 1, 1]$, 256 $\times 3$
Block Conv2	$[3, 3, 3]$, 128 $[3, 3, 3]$, 128 $\times 2$	$[3, 3, 3]$, 128 $[3, 3, 3]$, 128 $\times 4$	$[1, 1, 1]$, 128 $[3, 3, 3]$, 128 $[1, 1, 1]$, 512 $\times 4$
Block Conv3	$[3, 3, 3]$, 256 $[3, 3, 3]$, 256 $\times 2$	$[3, 3, 3]$, 256 $[3, 3, 3]$, 256 $\times 6$	$[1, 1, 1]$, 256 $[3, 3, 3]$, 256 $[1, 1, 1]$, 1024 $\times 6$
Block Conv4	$[3, 3, 3]$, 512 $[3, 3, 3]$, 512 $\times 2$	$[3, 3, 3]$, 512 $[3, 3, 3]$, 512 $\times 3$	$[1, 1, 1]$, 512 $[3, 3, 3]$, 512 $[1, 1, 1]$, 2048 $\times 3$
avg pool	output size $1 \times 1 \times 1$		
Dropout	rate 0.3		
Dense			

5.1.2. Model Training

We now pass on the considerations of the training phase. The model weights are initialized using the He/Kaiming initialization [7], which is specifically designed for networks that use rectified linear activation functions (ReLU). This method helps prevent the signal from vanishing or exploding as it propagates through the layers, setting the initial weights of each neuron in a layer by sampling from a Gaussian distribution with zero mean and a variance of $\frac{2}{n}$, where n is the number of inputs to the neuron.
As discussed in the data processing section, even after the pre-processing, the data results are still unbalanced. To mitigate this problem, class weights are computed based on the label frequency. These weights are used directly into the loss function which is a variation of the *Categorical Cross-Entropy*, The *Focal Loss* [8].
The focal loss aims to alleviate the problem of unbalanced data, by introducing a modulating factor that reduces the loss assigned to well-classified examples and focuses

more on the hard or misclassified examples.
Moreover, in each set of frames from the dataset are extracted only 60 frames, passing from a video sequence of 30 fps to 15 fps. The first frame of the trained sequence is $F_0 \sim \text{Uniform}(0, 180)$ since the total frames for each clip are 300 and we half the sampling frequency.
To take into account the problem of overfitting, besides the dropout layer and batch normalization in the model, other techniques are exploited: early stopping, L2 regularization with weight decay equal to $1e^{-3}$.
The early stopping method monitors the performance of the model on a validation set, and stops the learning when the performances are not improving. Additionally, a patience value is included, as a new learning parameter, to avoid non-necessary stops.
Further learning details are included such as:

- *Adam optimizer*, an adaptive optimization algorithm commonly used in training deep learning models.
- *Learning Rate scheduler*, used in training machine learning models to adjust the learning rate during the training process. In the specific the one cycle scheduler [9] with a max learning rate of $1e-4$.
- *Gradient scaler*, to reduce the range of magnitudes in the gradients.

Other learning parameters vary based on the experimental setup, i.e. batch size, epochs, patience, etc.
We conclude this section by leaving significant data on the training time, necessary for two versions of the custom dataset. Using the whole collection of videos from the enhanced version (v1), the training time for a single epoch is about 4h. Instead, using the reduced version with face extraction (v3), the time required for the same operation is around $2'\ 30''$,

5.2. Gaze Pipeline

The second workflow that we present, aims to extract attention information from the gaze of the interlocutor. The direction and focus of the gaze offer valuable cues about where individuals allocate their visual attention and the objects or regions that capture their interest. When individuals are attentive to a specific task or stimulus, their eyes tend to focus on relevant objects or regions of interest. To accomplish this task, different computer vision neural networks and techniques have been applied to obtain coherent information on the gaze target. In the specific, two models are involved from the *dlib* [10] library.

5.2.1. Features extractor

The first process consists uses a pre-trained CNN specifically designed for face detection. This model is trained to detect faces in images and provides bounding box coordinates for each detected face. It is based on the Histogram of Oriented Gradients features and a linear Support Vector Machine classifier.
The following step uses a pre-trained shape predictor model for facial landmark detection. This specific model is trained to detect 68 facial landmarks on a face.
Facial landmarks refer to specific points on the face, such as the corners of the eyes, the tip of the nose, and the corners of the mouth.
Thanks to these two first operations performed by the features extractor, a full set of face key points is obtained, in addition to the face patch and the eyes' patches of the frame (both converted to grayscale).

5.2.2. Analyzer

The features extracted are subsequently used from the analyzer module, which is in charge of providing two fundamental information, the face orientation in reference to the vertical axis (yaw angle), and an estimation of the gaze direction.
The orientation is computed considering the angle of the vector that starts from the eyes' midpoint and reaches the nose tip. To make this calculus robust also to longitudinal rotations (roll angle), this vector is projected to be always orthogonal to the horizontal axis when the interlocutor is directed toward the camera.
For the estimation of the gaze direction instead, starting from the eyes' patches and eyes' landmarks, is first applied a mask to remove pixels that not belongs to the eyes. Then, the patches are binary segmented using the Otsu's method [11], which automatically calculates an optimal threshold value for image segmentation. This thresholding technique can automatically adapt to varying lighting conditions and uneven illumination, making it particularly useful in situations where manual threshold selection may be challenging or time-consuming, improving the robustness.
The segmented patches are then partitioned horizontally to study the horizontal gaze direction, and vertically for the vertical gaze direction.An equivalent scheme of these steps is provided in figure 2.
Studying the distribution of the white pixels of these partitions with respect to the total number of active pixels, subtracting the percentual quantities over opposite directions (left-right and up-down) gives a discriminant on the general direction.
For instance, a higher density of white pixels towards the upper region suggests an upward gaze tendency, while a higher density towards the lower region indicates a downward gaze tendency. Similar interpretations can be made for rightward and leftward gaze orientations.
Therefore, two different relative values are calculated,

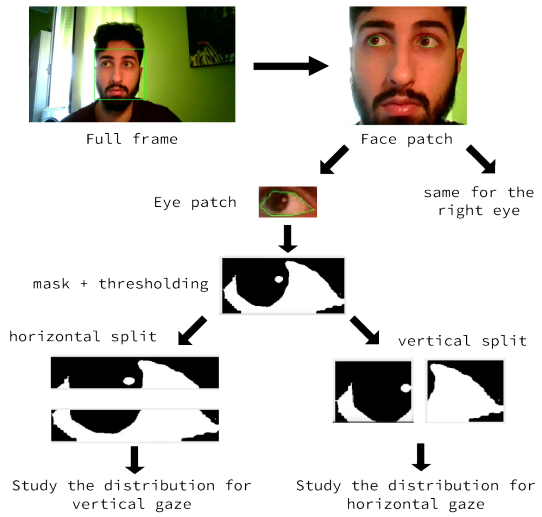


Figure 2: Gaze inference

for both horizontal and vertical directions. The analysis of white pixel distribution provides us essential features for gaze estimation. By studying the spatial distribution, we can identify the dominant areas of interest, such as the center or corners of the region. Additionally, the density of white pixels within these areas can offer valuable clues about gaze direction.

5.2.3. Gaze Scorer

To capture the changes in gaze direction, it is crucial to define limits that reflect the range of permissible shifts in specific directions. They are characterized by a set of 4 values, one for each cardinal direction.
These limits can be predefined based on the characteristics of the task or context. For example, in a driving scenario, the limits for vertical gaze shifts may be determined by the road and dashboard layout. Extracting these limits ensures that the gaze estimation model considers the context-specific constraints when assessing the magnitude of gaze changes. In this case, we refer to limits in the context of an interaction between a person and a device endowed with a camera, like a computer.
To make the system robust, these limits should be dynamic, in the sense that can change based on the user state.
In the application proposed the limits change accordingly to the face orientation and the face position with respect to the camera.
The gaze scorer module is in charge of updating these limits if necessary and computing vertical and horizontal errors that reduce the gaze attention estimation.
Whether face orientation is greater than 20 degrees (absolute angle) with respect to the frontal position, horizontal limits are shifted by the sine of the face orientation.
Instead, for what concern update based on face position, is computed the distance between the face box center and the frame center, if this exceeds the sixth of the frame dimension, limits are updated. The errors are computed with respect to the actual gaze estimation and the limits, then these values are normalized. Finally, the gaze score is obtained by subtracting these errors from the maximum score (1) and clipping to acquire a range between 0 and 1. This scorer will return a value of 0 in the case that no face has been detected in the frame.

5.3. Attention Scorer

To obtain the final output, the two pipelines are seamlessly integrated. The engagement pipeline's predictions and the gaze-related attention pipeline's results are fused and combined using a suitable fusion strategy.
This integration allows us to leverage the strengths of both pipelines and capture a coherent understanding of the interplay between engagement and gaze-related attention.
The first operation to perform is the conversion from engagement prediction labels to an engagement score. This is conducted following the intuitive relation expressed in the table 4.

Table 4
Relation table between engagement degree label and score

Label	*Score*
0	0.1
1	0.4
2	0.7
3	1.

The final score is obtained using a simple weighted linear interpolation formula of this type:

$$\alpha \cdot score_{gaze} + (1 - \alpha) \cdot score_{eng}$$

We conclude this section by summarizing the overall system presented through the figure 3.

6. Results

After the commentary on the implementation details of the system, we introduce here the description of the experimental setup and quantitative results.
The evaluation of engagement degree in a multi-class problem poses unique challenges due to the subjective

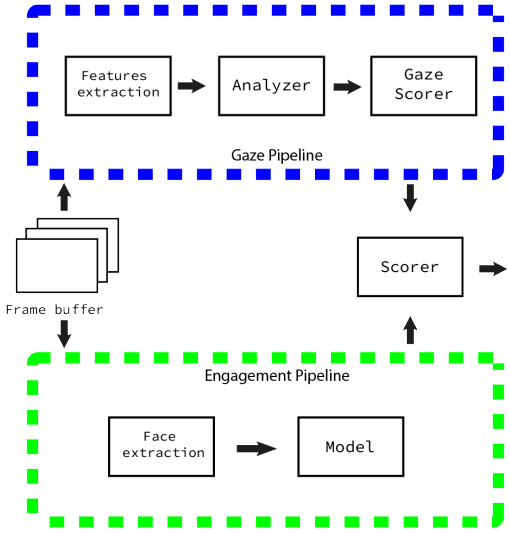


Figure 3: System architecture

and continuous nature of engagement levels. While traditional classification metrics can be informative for binary or categorical problems, they may not fully capture the nuanced variations in engagement that exist across different individuals or scenarios.
To address this, we adopt a regression-based evaluation approach that uses both scores from the overall system and the relative measures from each engagement level in the test set, allowing for a more granular assessment of user engagement. Once more the conversion from label to score and the opposite is performed following the relationship in table 4.
By using regression metrics, such as mean absolute error (MAE) and Mean absolute percentage error (MAPE), we can quantify the discrepancy between the predicted engagement scores and the ground truth values. This allows us to evaluate the performance of our model in terms of its ability to estimate engagement and attention levels accurately. Additionally, regression metrics provide a more continuous and interpretable measure of the model's predictive capability, enabling us to gauge the extent to which it captures the subtle variations in engagement across different contexts.
it's crucial also to notice that such metrics tend to penalize more misclassified engagement labels whether the expected values and the predicted values represent two distant degrees.
Measuring the attention level using engagement labels is a valid approach because engagement and attention are closely related concepts. While engagement encompasses a broader range of cognitive and emotional factors, attention is a fundamental component of engagement.

When measuring the attention level, we can leverage the engagement labels as an indirect measure of attention.
For completeness, it's shown how these metrics changes in reference to different values of the interpolation parameter α of the scorer.
Before doing this we expose the data about Hyperparameters and other information about the model and dataset in the table 5.

Training parameters	*Value*
Batch size	16
ResNet version	34-layer
Dataset version	v3
Epochs	50
Patience	10

Table 5
Engagement model setup

The dataset version used for learning is the v3, which allows faster and more accurate training. All the other learning details are the ones already exposed in the model section.
Whereas the entire set from v1 is used for testing the complete application, in order to have an evaluation that takes into account the maximum possible number of instances.
The accuracy gained from only the engagement model is close to 50% (49.7%), which is a similar result compared to performance acquired in previous works, such as the one presented forward by Liao et al. [12].
the results achieved using regression metrics to evaluate the whole application are presented in figure 4 and figure 5.

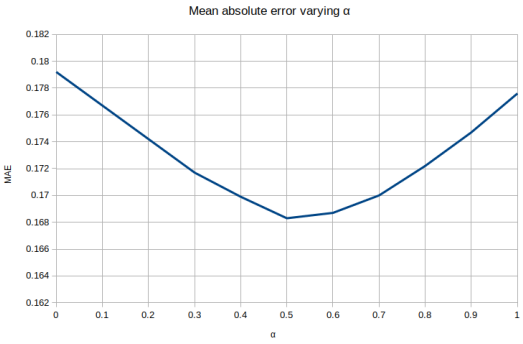


Figure 4: MAE metric changing over different values of α

Besides all the improvements carried out in the data transformation step, reducing consistently the amount of required storage space, and speeding up the training phase by about 96 times, we can be satisfied with

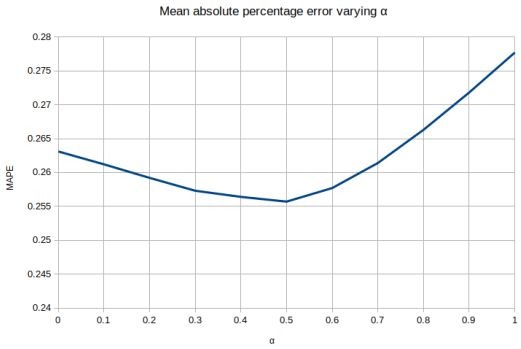
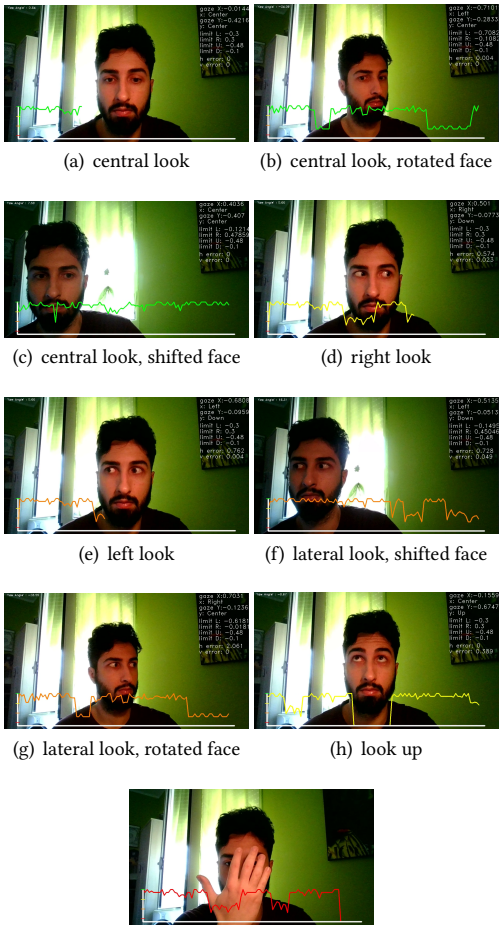


Figure 5: MAPE metric changing over different values of α

the limited error achieved by the system in this test. Especially considering the subjective nature of the engagement level estimation.
Moreover, it's clear how a balancement of the output from the two workflows, can lead to have a benefit in terms of results. In fact, the best metrics are recorded for $\alpha = 0.5$, improving the inference from just using gaze pipeline ($\alpha = 1$) or the engagement one ($\alpha = 0$).
The absence of a reliable baseline makes it difficult to evaluate the effectiveness and performance of our proposed approach in a comparative manner. Without a benchmark to compare against, it becomes challenging to determine the extent to which our method outperforms or deviates from existing techniques.

7. The Real-Time application

Afterward the illustration of data coming from the experiments performed, we want to give some qualitative outcomes and expose the real-time application that implements the discussed system.
The application makes use of all the components discussed in the methodology section, including also a webcam reader module, to elaborate the webcam frames in real-time, and a plotter, to draw the attention score on a chart.
The chart, as shown in the figure 6, uses the x-axis for the time (frames), and the y-axis for the attention level.
The last value on the plot represents the actual attention value, and the trend is colored based on the attention label's bounds defined in the table 4.
Moreover, the system needs an initial time to fill a buffer of 60 frames necessary for the engagement model. Then, this buffer is updated each time with the new frame coming from the webcam reader module.



(a) central look (b) central look, rotated face

(c) central look, shifted face (d) right look

(e) left look (f) lateral look, shifted face

(g) lateral look, rotated face (h) look up

(i) Face occlusion

Figure 6: Multiple scenarios in the application

8. Conclusions

In this paper, we presented an attention estimation system that utilizes computer vision techniques to infer the attention level of individuals. Through the integration of various components such as gaze estimation, facial analysis, and engagement recognition, our system offers valuable insights into understanding human attention in real-world scenarios.
Our research focused on addressing the challenging task of estimating attention in dynamic and uncontrolled environments. By leveraging the power of machine learning, we were able to develop a model capable of estimating in an appreciable manner the engagement degree based on a combination of spatial and temporal information.
One of the key contributions of our work lies in the

incorporation of gaze information as a crucial factor for attention estimation. By analyzing the distribution of white pixels in segmented gaze regions, we obtained valuable insights into the direction of gaze, enabling us to infer in which direction an individual is looking. This information proved to be valid in capturing changes in attention and enhancing the overall accuracy of our system.

In conclusion, our work represents a novel approach in the realm of attention estimation, combining several methodologies, extensive data processing, and an intuitive real-time application, we have established the possibility for further exploration and innovation in understanding and evaluation of human attention.

8.1. Future work

While our attention estimation system has achieved promising results, there are several avenues for future exploration and enhancement. Here are outlined potential directions for future work that can further advance the field of attention analysis and its practical applications.

Our current system focuses on estimating attention for a single individual. An interesting extension would be to enable attention estimation in multi-face contexts, where multiple individuals are present simultaneously, so pass to a Multi-face Attention Estimation.

Another useful possible work would be the improvement of the engagement model, incorporating data from a more diverse and heterogeneous group of people. This would involve collecting data from individuals with varying demographic characteristics, cultural backgrounds, and engagement patterns. By including a broader range of individuals, we can enhance the model's robustness and generalizability.

A different related improvement could be also the integration of additional data sources. Our attention estimation system exploits visual information obtained from video streams. However, there are other potential data sources that can enrich the analysis of attention. For instance, integrating data from stereo camera systems can provide depth information and enable more accurate depth-based attention estimation. Incorporating audio data, such as speech patterns and intonation, can also provide valuable cues for understanding attention. Exploring the fusion of multiple modalities can lead to a more comprehensive analysis of attention.

By addressing these possible paths for future work, or other useful improvement strategies, we can continue to push the boundaries of attention analysis in order to improve the various practical applications.

References

[1] F. D. Duchetto, P. Baxter, M. Hanheide, Are you still with me? continuous engagement assessment from a robot's point of view, CoRR abs/2001.03515 (2020). URL: https://arxiv.org/abs/2001.03515. arXiv:2001.03515.

[2] 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.

[3] 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 (2022) 1800–1808. doi:10.1109/TIE.2021.3057033.

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

[5] E. Cengil, A. Çınar, E. Özbay, Image classification with caffe deep learning framework, in: 2017 International Conference on Computer Science and Engineering (UBMK), 2017, pp. 440–444. doi:10.1109/UBMK.2017.8093433.

[6] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, CoRR abs/1512.03385 (2015). URL: http://arxiv.org/abs/1512.03385. arXiv:1512.03385.

[7] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, CoRR abs/1502.01852 (2015). URL: http://arxiv.org/abs/1502.01852. arXiv:1502.01852.

[8] T. Lin, P. Goyal, R. B. Girshick, K. He, P. Dollár, Focal loss for dense object detection, CoRR abs/1708.02002 (2017). URL: http://arxiv.org/abs/1708.02002. arXiv:1708.02002.

[9] L. N. Smith, N. Topin, Super-convergence: Very fast training of neural networks using large learning rates, 2018. arXiv:1708.07120.

[10] D. E. King, Dlib-ml: A machine learning toolkit, Journal of Machine Learning Research 10 (2009) 1755–1758.

[11] N. Otsu, A threshold selection method from gray-level histograms, IEEE Transactions on Systems, Man, and Cybernetics 9 (1979) 62–66. doi:10.1109/TSMC.1979.4310076.

[12] J. Liao, Y. Liang, J. Pan, Deep facial spatiotemporal network for engagement prediction in online learning, Applied Intelligence 51 (2021) 6609–6621.