Microscopic Binary Engagement Model

Marco Lemos[0009−0004−8727−4254], Pedro J.S. Cardoso[0000−0003−4803−7964], and João M.F. Rodrigues[0000−0002−3562−6025]

NOVA LINCS & ISE, Universidade do Algarve, Faro, Portugal
{a72178,pcardoso,jrodrig}@ualg.pt

Abstract. Tracking audience engagement in real-time offers numerous benefits. For instance, event planners can make dynamic adjustments to presentations or activities to maintain high levels of interest and participation. This enhances the overall experience for attendees by ensuring the content remains engaging and relevant. This paper proposes a model for computing the binary engagement within groups. The model does this by identifying individuals' engagement during the events' time frames, which are then combined, i.e., the engagement of the group is computed by aggregating the engagement of each individual. For each individual of the group, the engagement model incorporates the computation over time of the gaze direction, valence, and arousal, classifying the engagement into two primary levels: not-engaged and engaged. The engaged category is further divided into two sublevels: positive and negative engagement. Experimental results confirm the model's effectiveness, showcasing reliable identity tracking and accurate assessment of engagement states in dynamic scenarios.

Keywords: Engagement · Affective Computing · HCI · Group Engagement · Real-time Engagement.

1 Introduction

Detecting audience engagement in real-time during events is a cutting-edge approach that leverages advanced technologies to measure and analyze how attendees interact and respond throughout an event. This process may involve various data collection methods, such as video analysis, audio cues, physiological sensors, and social media monitoring, to capture real-time feedback.

Machine learning and computer vision advancements have paved the way for engagement understanding by combining human emotions through automated analysis of visual and behavioral cues. Emotional states, many times represented in a valence-arousal space (e.g., [3]), provide valuable insights into individual and group behaviors. Likewise, engagement levels (e.g. [15, 24]) offer an understanding of a person or group focus and participation in activities.

Although several studies emphasize personal or group engagement online, such as social media engagement [5, 8] or learner engagement with virtual educational events [4], in the context of real-world and real-time (live) engagement detection during indoor and outdoor events, very few models have been proposed [15].

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_9

At this point, it is important to define the terms crowd and group [13]. A *group* is a collection of individuals, ranging in size from two to hundreds, who are present together at any given time and engaging in social contact. Its members move in a similar direction and at a similar speed, making them near to one another. Multiple groups can cohabit during an event. Conversely, a *crowd* (or mass) is a special huge gathering of people who are physically present in the same place. It typically arises when individuals who have a common objective unite as a single entity, losing their individuality and assuming the characteristics of the crowd entity.

The engagement analysis in groups and crowds can be divided into two main methodologies [15, 20]: *Microscopical* (or bottom-up) methods, typically applied to groups, where individuals in the "video streaming" are analyzed and the resulting data is then used to extrapolate information at the collective level, i.e., a group analysis is considered as a collection of individuals analysis; *Macroscopical* (or top-down) methods, typically applied to crowds, are made up of comprehensive processes that view the crowd as a single cohesive unit, rather than requiring the tracking and segmenting of every individual. Macroscopic approaches are (more) suited when population density increases and tracking quality drastically decreases.

It is also important to define *instantaneous engagement*, which corresponds to the engagement detected at each instant t (or frame f) of the stream or video. Similarly, the *period engagement* corresponds to the engagement for a specific time period, while *event engagement* accounts for the engagement throughout the entire event. For more details see [15] and Section 3.

This paper focuses on engagement in groups, i.e., when there are few or no occlusions, low density, and a clear view of people. In such cases, microscopic approaches frequently perform best. By using a microscopical approach, this paper focuses on presenting a modular and scalable model for instantaneous, period, and event engagement detection in groups during real-world events – Microscopic Binary Engagement Model (MiBE). The model integrates person tracking, with the combination of two dimensions: (i) emotion (valence and arousal level estimation) and (ii) attention (focus-head pose estimation). More dimensions can be integrated in the future [15].

The main contribution of the paper is a scalable binary engagement detection model for groups, where engagement can be classified as positive if the person is "appreciating/liking" the event, or negative if the person despite being engaged, is not "appreciating/liking" the event. A secondary contribution is the introduction of an initial model for valence-arousal computation.

In the present section, the subject and goals of the paper are presented. Section 2 briefly summarizes the state of the art. Section 3 introduces the proposed model – MiBE, and Section 4 presents the initial tests and results achieved. The final section outlines some conclusions and future work.

2 Related work

As already mentioned, MiBE is based in two dimensions: emotion and attention. Here we will not go into detail on the different emotion and attention models,

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_9

we will rather briefly enumerate some recent models. For emotion computation the valence-arousal (VA) predictions can be used. Valence is a measure of the emotional intensity, ranging from negative to positive, while arousal indicates the emotional intensity, ranging from low to high. Nguyen et al. [14] present an approach to affective behavior analysis, focusing on VA prediction within the Affective Behavior Analysis in the Wild (ABAW3) challenge. Leveraging deep learning (DL) techniques, the authors propose a two-stage model for continuous emotion estimation. Experimental results on the Aff-Wild2 dataset demonstrate significant improvements over baseline methods, achieving a Concordance Correlation Coefficient core of 0.507 for VA estimation and an F1-score of 0.533 for action unit detection. Stephen et al. [11] presented a DL method for predicting continuous affect from facial expressions (FE) in the VA space. The method maps discrete emotion labels and FE to this space, outperforming existing methods on the AffectNet dataset [13] and showing strong generalization. Andrew [16] introduced a real-time video-based algorithm for predicting FE, VA, and action units on mobile devices. Lorenzo et al. [1] explore VA estimation from neuromorphic vision data using event cameras, which excel at capturing subtle and rapid facial micro-movements. Other models also exist, such as the one proposed in [3].

For attention detection, head pose estimation (HPE) can be used. The Wide Headpose Estimation Network (WHENet) [25] is a model designed for HPE using single RGB images. It excels in predicting Euler angles—yaw, pitch, and roll—over a full 360-degree yaw range, which is critical for applications like autonomous driving and augmented reality. Built on the EfficientNet-B0 backbone, WHENet combines regression and classification objectives for robust and fine-grained pose prediction. Evaluation on BIWI and AFLW2000 datasets shows WHENet achieving a mean absolute error as low as 3.81 degrees. Hempel et al. [9] introduce a method from single images using a continuous 6D rotation matrix representation. Later, the same authors used a geodesic loss function within the Special Orthogonal Group to stabilize learning and ensure precise predictions [10]. The model, named 6DRepNet360, is open-sourced to facilitate further research and application development.

Finally, there are models designed to detect engagement. Gupta et al. [7] introduce a real-time DL-based learner engagement detection system that leverages facial emotion recognition (FER). Addressing the challenges of online education, it measures student engagement by analyzing facial expressions captured via webcams during online sessions. Lasri et al. [12] detect the engagement levels of deaf and hard-of-hearing students through FER. More recently, Zhao et al. [24] present a model designed to detect student engagement through FE in real-time classroom settings.

According to the literature that has been presented and examined, no model has been found that can handle actual events that take place both indoors and outdoors and that can aggregate the engagement of various cameras, groups, and time periods; in other words, it cannot drill down the information from the individual's engagement with each object to the group, to the period, to the entire event.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_9

3 Binary Engagement Model

Before going into the detail of the model, let us define C as the combined information from different dimensions, $D_1, D_2, \ldots, D_n$, where n is the number of dimensions. A dimension refers to "emotion", "sentiment", "scene dynamics", "attention" etc. (for further details see [15]). With this in mind, let us also define *instantaneous engagement* as $IE(t, G) = C\{D_1, ..., D_n\}$, which corresponds to the engagement detected at time t (or frame f) of the streaming/movie for a non-empty set G of individuals. If G is a set with a single individual, $G = \{h\}$, then it will be the engagement of the person h at time t. If G is a set with more than one individual, $G = \{h_1, h_2, \ldots, h_n\}$, then it will be the engagement of the group at time t. The *P-period engagement* is given by $PE(P, G) = \uplus_{t \in P} IE(t, G)$, where $P = \{t_i, t_{i+1}, \ldots, t_f\}$ is a period of time, t_i and t_f (with $t_i < t_f$) are two different times in the event timeline, and $\uplus$ is the combination of the information retrieved from the different instants. Finally, the *event engagement* is given by $E(G) = PE(I, G) = \uplus_{t \in I} IE(t, G)$, i.e., it accounts for the entire event, with duration interval I.

In the present model, the $P-$period engagement for a group (G) is computed as the mean engagement of all persons in the group, i.e., $PE(P, G) = \frac{1}{|G||P|} \sum_{h \in G} \sum_{t \in P} IE(t, \{h\})$, where $|.|$ is the number of elements of the set. Similarly, the $E(G)$ is computed as the mean engagement of all persons in the group, i.e., $E(G) = \frac{1}{|G||I|} \sum_{h \in G} \sum_{t \in I} IE(t, \{h\})$. Both $PE(P, G)$ and $E(G)$ can be computed for a single person h by setting $G = \{h\}$.

Furthermore, we defined binary levels of engagement that are determined based on valence, arousal, and gaze direction, whether the person is looking or not to the point of interest/scene (PoI). These levels are represented as pairs, $IE = [x, y]$, of binary (0-1) values, as follows. (i) **Not Engaged ($IE = [0, \times]$)** is distinguishable in (i.1) $IE = [0, 0]$ if the person is not looking at the PoI, regardless of their valence and arousal; or (i.2) $IE = [0, 1]$ if the person display a negative arousal (low emotional intensity) but is looking at the PoI. The latter suggests that the person is disinterested in the activity and not engaged, although their gaze being directed at the PoI. (ii) **Engaged ($IE = [1, \times]$)** which encompasses all other situations, divided in: (ii.1) **Negative Engaged**, $IE = [1, 0]$, if a person has a negative valence (expressing a negative sentiment) but a positive arousal (indicating a strong emotional intensity), and are looking (gaze) at the PoI. In this case, even though the person feels negatively about the activity, the high arousal indicates its engagement or that the reaction to the activity is strong. (ii.2) **Positive/True Engaged**, $IE = [1, 1]$, if a person displays a positive valence (expressing a positive sentiment) and a positive arousal (indicating a strong emotional intensity) while looking (gaze) directly at the PoI. This combination suggests that the person is actively and positively engaged with the activity. Therefore, the instantaneous engagement $IE(t, G)$ will be equal to 1 if the person is engaged negatively or positively, and 0 if the person is not engaged at all.

The model's global block diagram is shown in Fig. 1, and operates as follows. The first block is (a) *Head Box Detection*, which identifies bounding boxes

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_9

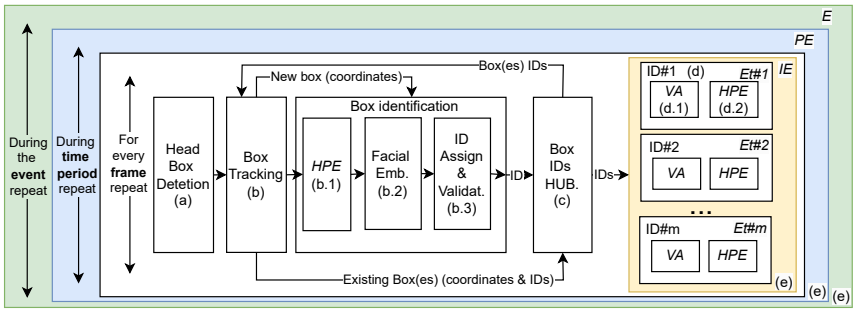

Fig. 1. Block Diagram of the MiBE model.

corresponding to heads in each frame. Next, the model performs (b) *Box Tracking*, where it tracks the previous identified boxes. For all boxes that matches an existing ID, the ID and box coordinates are passed to step (c). For any new bounding boxes (i.e., new heads), the (b) *Box Identification* is initiated and computes the (b.1) *HPE*, to determine head orientation. Then (b.2) *Facial Embeddings* are computed, to check if the face within the box resembles one from a previous frame. Finally, in step (b.3) *Box ID Assign & Validate*, the face is either assigned a new unique identity or an existing ID is validated and maintained.

Next, the (c) *Box ID HUB* is performed for both existing and newly processed bounding boxes, functioning as a central hub for managing and updating the box IDs. (d) For each of these bounding boxes (and for each dimension, $ID\#1$ to $ID\#n$) the model computes: (d.1) valence-arousal estimation to measure the individual's emotional state and (d.2) HPE for head orientation, computed once *per* frame (used also in step (b.1)) for each specific ID. (e) These computed dimensions are then used to calculate the values of engagement of the individuals and groups ($IE(t, G), PE(P, G)$, and $E(G)$), incorporating information from all groups involved, if more than one group exists.

Before going in details with each mentioned block, let us explain in more detail the HPE and VA estimation blocks.

3.1 Head Pose Estimation Block (HPE)

For each frame, the HPE was computed using WHENet[1] model [25], which allows to predict *yaw*, *pitch*, and *roll* values. In this context, *yaw* (ψ) indicates the degree of head turn to the left or right, with positive values indicating a turn to the right and negative values indicating a turn to the left. The range of yaw is $\psi \in [-180°, 180°]$, with $\psi = 0°$ indicating that the person is looking directly at the camera. *Pitch* (θ) represents the degree of head tilt up or down, with positive values indicating a tilt up and negative values indicating a tilt down. The range of pitch is $\theta \in [-90°, 90°]$. *Roll* (φ) represents the degree of head tilt to the left or right, with positive values indicating a tilt to the right and negative values indicating a tilt to the left. Roll values range in $\varphi \in [-180°, 180°]$.

[1] Model available at: https://tinyurl.com/yc47z9w3, accessed on 2025/01/16.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_9

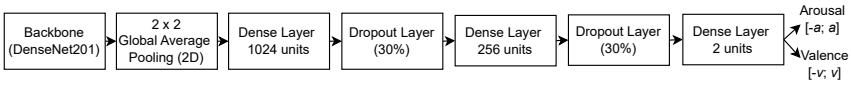

Fig. 2. Valence-Arousal model block diagram.

3.2 Valence-Arousal Estimation Block (VA)

The valence-arousal computation is performed concurrently with the HPE process through a newly developed VA Deep Neural Network (DNN) model - EVAm (see Fig. 2). The primary goal of the new model is to be seamlessly incorporated into the engagement framework, i.e., a model that can operate under real-world circumstances, using various cameras, positioned at various facial angles, in various lighting situations etc.

The initial model, still in its early steps, uses DenseNet201 pre-trained weights from ImageNet as the backbone of the DNN architecture. In the head of the DNN, the (i) first layer is a 2×2 *Global Average Pooling (2D)*, this choice preserved spatial information across channels while significantly reducing the number of trainable parameters compared to a Flatten operation [22]. This not only improved computational efficiency but also mitigated the risk of overfitting. The next layer (ii) is a *Dense Layer* with 1024 units (neurons), incorporated to learn complex feature interactions from the pooled features, followed by a (iii) *Dropout Layer*, to enhance regularization and prevent overfitting. The Dropout Layer, with a 30% dropout rate, is repeated after each dense layer in the architecture. The next layer (iv) is a *Dense Layer* with 256 neurons for further feature abstraction, and finally, the output (v) is a *Dense Layer* with 2 neurons used to predict respectively the Valence and Arousal values simultaneously.

Training, Results and Discussion (Partial Results) We used the AffectNet dataset [13] for training, validation, and testing. The dataset contains over 400,000 facial images that have been manually labeled for the presence of eight different facial expressions. Additionally, the dataset includes annotations for VA intensity. For training, we utilized 288,000 images annotated with valence and arousal. The validation and testing sets each comprised 2,000 images.

In the training phase, Early Stopping was introduced with a 10 epochs patience, to prevent overfitting and reduce training time by halting the process once performance stagnated. Learning Rate Scheduling was employed with a 5 epochs patience and a factor of 0.5, allowing the optimizer to reduce the learning rate when progress slowed. For the dense layers, the Rectified Linear Unit (ReLU) activation was employed. To optimize the model, we employed the Adam optimizer and a mean squared error (MSE) loss function, as it is well-suited for regression tasks, penalizing larger deviations more heavily. Training was conducted with a batch size of 64, balancing memory efficiency and gradient stability for effective optimization. Finally, the DenseNet-201 model architecture requires images to be of size 224×224 pixels.

To facilitate efficient batch processing and ensure consistent model performance, the ImageDataGenerator for both data preprocessing and augmentation

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_9

Model / Metrics	Valence					Arousal				
	MAE	RMSE	PCC	CCC	SAGR	MAE	RMSE	PCC	CCC	SAGR
EAVm (ours)	0.290	0.390	0.630	0.610	0.760	0.280	0.360	0.560	0.470	0.760
Mollahosse et al. [13]	-	0.370	0.660	0.600	0.740	-	0.410	0.540	0.340	0.650
Stephen et al. [11]	0.146	0.179	0.952	0.948	-	0.121	0.164	0.952	0.950	-
Andrey [16]	-	-	-	0.429	-	-	-	-	0.496	-

Table 1. Valence-Arousal model results, with MAE - Mean Absolute Error, RMSE - Root Mean Square Error, PCC - Pearson Correlation Coefficient, CCC - Concordance Correlation Coefficient, and SAGR - Sign Agreement Ratio.

was employed. A key preprocessing step involved normalizing pixel values to the range [0, 1] by dividing by 255. This normalization accelerates convergence and stabilizes the training process by standardizing the input, as highlighted in [6].

As for results, while our model presents good performance, see Tab. 1, it still falls short when compared to other models. When compared with the results in [13], the baseline, our model shows superior performance across all metrics for arousal. For valence, our model outperforms the model in two metrics, while the other two remain close to the baseline values. When comparing with the results in [11], the authors presented better results than us, but they only used images corresponding to seven of the eight existing emotions in the AffectNet dataset, they exclude the neutral emotion. In the case of Andrey [16], which uses the Aff-Wild2 dataset, our results demonstrate better performance in valence and similar results in arousal. However, since Andrey's work utilized a different dataset, a direct and fair comparison is not feasible.

Finally, it is important to stress that this is a first version of the model. Future work will involve testing different backbones and applying various approaches to the DNN head. Additionally, the model will be trained on a combination of datasets to have a better generalization.

We will now explain in more details the remaining blocks of the model.

3.3 Head detection, Tracking and Identification Blocks

Head Box Detection This is the first module where the streaming input is processed. The focus it to detect heads, rather than retain any facial image information. In compliance with the General Data Protection Regulation (GDPR), only the bounding boxes containing heads are of interest. This initial detection is performed using the YOLOv4 [2] model[2], trained on the Hollywood Heads [21] and CrowdHuman [19] datasets. More recent models exists but, for this initial prototype, YOLOv4 presented a good solution, with a good balance between accuracy and speed.

Box Tracking For each frame, in the *Box Tracking* module, the centroid of each bounding box is computed and compared with the centroid(s) of the previous

[2] Model available at: https://tinyurl.com/2pbydwn7, accessed on 2025/01/16

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_9

frame. The Euclidean distance between centroids is tested against a threshold defined as half the box's width (w_{box}), i.e., $th_{bt} = w_{box}/2$. If the distance is less than th_{bt} then (i) the box is considered the same as the one in the previous frame, and the box coordinates along with its corresponding ID are sent to the *Box ID HUB* module. If the distance is greater than th_{bt} then (ii) the box is considered as a new box, and the box coordinates are sent to the *Box Identification* module.

Box Identification In this module, the initial/new boxes are identified. To account for variations in head orientation, the *HPE* (see Sec. 3.1) is computed for each box. Faces within the boxes are "acknowledged" using the MTCNN model [23] from the DeepFace library [18] and each box is then processed by the FaceNet model [17], which computes facial embeddings – vector representations that capture the unique features of the face – *Facial Embedding* module.

Any new boxes detected in subsequent frames are processed by computing embeddings with FaceNet. These embeddings are compared to those in the stored database using cosine similarity. If the similarity score between a new embedding and the stored embeddings exceeds a predefined threshold (0.4, given by the FaceNet model), the system at *ID Assign & Validation* module assigns the box to an existing ID. Otherwise, a new ID is generated. This approach ensures consistent recognition of individuals, even when they temporarily leave and re-enter the scene, ensuring that no image is stored.

By using the HPE information, as the head orientation changes, the system computes multiple embeddings for each individual, representing different angles such as *yaw* and *pitch*, ensuring the embeddings remain robust against changes in orientation, lighting, and facial expressions. By storing multiple embeddings per box (person), the system reduces mismatches caused by these variables.

Box ID HUB This module works like a Hub, receiving IDs and boxes coordinates from *Box Tracking* and from *Box Identification*, routing that information back to the *Box Tracking* module for the new position of the box, and at the same time to the *Instantaneous Engagement – IE* module, to the respective $IE\#1, \ldots, IE\#m$ (sub-)processes that are directly related with each box ID ($ID\#1, \ldots, ID\#m$).

3.4 Instantaneous Engagement (IE)

The next step is to compute the instantaneous engagement for each ID and frame ($IE(t, \{h_{ID\#i}\}), i = \{1, \ldots, m\}$), which are then combined to compute the G group instantaneous engagement $IE(t, G)$ (see Sec. 3).

First, let us define the Gaze (Ga) in relation to a PoI. The model operates, at the moment, under the assumption of static PoIs, a constraint dictated by the present design of our mapping approach. Initially, we have to generate a map of the room, setting the (x, y) coordinates of the individuals, cameras, and PoI. Using this map, we compute the angles between each person, the PoI, and the camera, to determine whether a person is looking at a specific PoI or not. Then, we compare those angles with the *yaw* values predicted by the HPE model (see

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_9

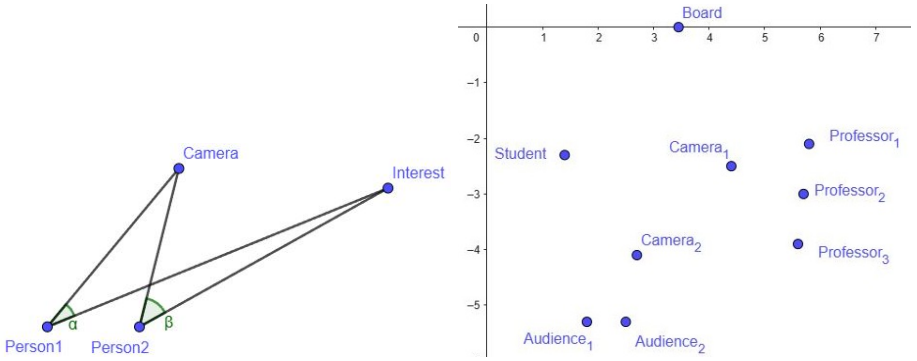

Fig. 3. On the left, a sketch of the angles between persons, PoI, and a camera. On the right, a room setup (see details in Sec. 4).

Sec. 3.1). To account for minor positional variations (e.g., leaning or the size of the PoI), we accept a ±25 cm (empirically chosen) shift of the coordinates of the persons, the cameras and the PoI, calculating maximum and minimum angles accordingly. This ensures a robust determination of gaze despite small positional shifts.

Figure 3 (left) illustrates an example of a mapping for two persons, a PoI, and a camera positioned within a room. The angles α and β, for each individual, are computed using standard trigonometric formulas. For instance, for $Person_1$, if d_p be the distance between $Person_1$ and the Camera, d_i the distance between $Person_1$ and the PoI, and d_c the distance between the Camera and the PoI then $\alpha = \arccos\left((d_i^2 + d_p^2 - d_c^2)/(2 \cdot d_i \cdot d_p)\right)$. Figure 3 (right) illustrates a real setup for a student presentation (more details are presented in Sec. 4).

If the predicted *yaw* and *pitch* angles fall within the calculated range, the person is classified as looking at the point of interest, $Ga = 1$. For example, considering Person1, the *yaw* angle, ψ_{p1}, returned by HPE, must lie within the interval $\psi_{p1} \in [\alpha - 10°\alpha + 10°]$. Additionally, the *pitch* value must be greater than $-10°$ to ensure the person is looking at the PoI. If the *yaw* angle lies outside the range or the *pitch* value is less than $-10°$, the person is deemed not to be looking at the PoI, $Ga = -1$.

It is important to note that the *yaw* value of 0° represents alignment with the camera's optical axis, while the *pitch* value of 0° reflects no upward or downward tilt of the head, independent of the camera's vertical alignment. This distinction becomes crucial in future scenarios where the PoI change its elevation, as our model does not currently account for such vertical movements. In addition, this model only considers the gaze direction based on the orientation of the head, without accounting for eye movements. In other words, it assumes that the eyes are looking straight ahead and aligned with the head's direction. This simplification has not yet been addressed in the current implementation.

Now, let us define the engagement for each frame (f) in a bi-dimensional space, namely as: $E_{level,\pm} = (Ga.(A+1)/2, V)$, where Ga is the gaze (computed as presented above), A is the arousal, and V the valence (A and V are

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_9

estimated with the model presented in Sec. 3.1). The first coordinate shows the level of engagement and the second establishes if the engagement was generated by a positive/"good" or a negative/"bad" emotion. In the formula, the arousal is normalized to the interval $[0, 1]$, before being multiplied by the gaze. This normalization step means that if the arousal is negative, the resulting value will be lower than if the arousal were positive. Consequently, for a negative arousal value, the engagement level - calculated by multiplying the normalized arousal by gaze - will be smaller than for a positive arousal value. The multiplication reflects how gaze intensity and arousal level together determine the overall engagement level. The second coordinate, representing valence, indicates the person's emotional state, i.e., a positive valence indicates that the person experiences positive engagement whereas a negative valence indicates negative engagement.

Thus far, the instantaneous engagement of each individual has been computed as, $IE(t, \{h_i\})$, where $i = \{1, \ldots, m\}$. As defined in Sec. 3, the instantaneous engagement of the group is computed as the mean instantaneous engagement of all individuals in the group, i.e., $IE(t, G) = \frac{1}{m} \sum_{i=1}^{m} IE(t, \{h_i\})$.

3.5 Period Engagement (PE) and Event Engagement (E)

Following the above, repeating the instantaneous engagement module for each frame, and using the formulae presented in the beginning of this section, it is now possible to compute the group's P-period engagement $PE(P, \{h_i\})$, $i = \{1, \ldots, m\}$ (for each person) and event engagement $E(\{h_i\})$, $i = \{1, \ldots, m\}$).

Using the same reasoning, it can be computed the engagement in the 4 binary segments presented initially: $[0, 0]$ - no engagement, $[0, 1]$ - disinterested, $[1, 0]$ - negative engagement, and $[1, 1]$ - true engagement.

It is important to stress, the MiBE is completely scalable in terms of individuals and groups, and can cope with the information of more than 1 camera.

4 MiBE Operational Tests and Assessment

To illustrate the functionality of the MiBE, we present two tests. In **Test#1**, the simplest setup is considered, involving only one person and a PoI which is the same as the camera positioned directly in front of the person. Figure 4 (left) shows a frame extracted from the video, with the person looking directly at the camera, resulting in $Ga = 1$. The person exhibits negative valence and arousal, indicating a negative emotional response toward the scene. The middle plot shows the valence and arousal values per frame (each frame is represented by a point in the VA plane), while the right plot shows the instantaneous engagement, $IE(t, \{h\})$, during the full stream.

The right plot reveals that the individual is not engaged during certain frames, which corresponds to moments when he is not looking at the PoI (camera). Additionally, when engaged, the person's engagement is negative, as indicated by the negative valence and low arousal values. Finally, $E(\{h\}) = 80\%$ for the length of the video, which has 6 seconds.

The second test – **Test#2** – was done in real indoor environment, during a student's master thesis presentation. Figure 3 (right) illustrates the layout setup

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_9

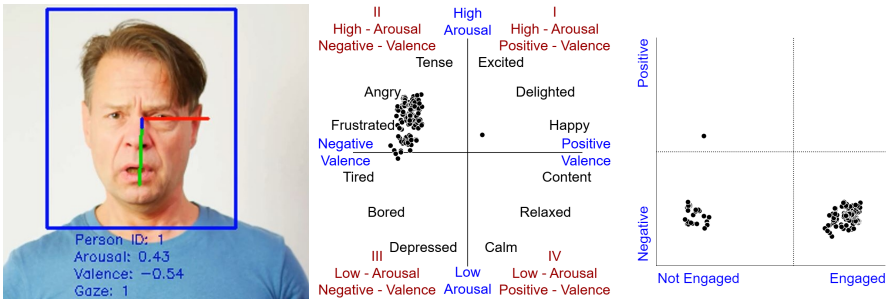

Fig. 4. Illustration of Test#1, see text.

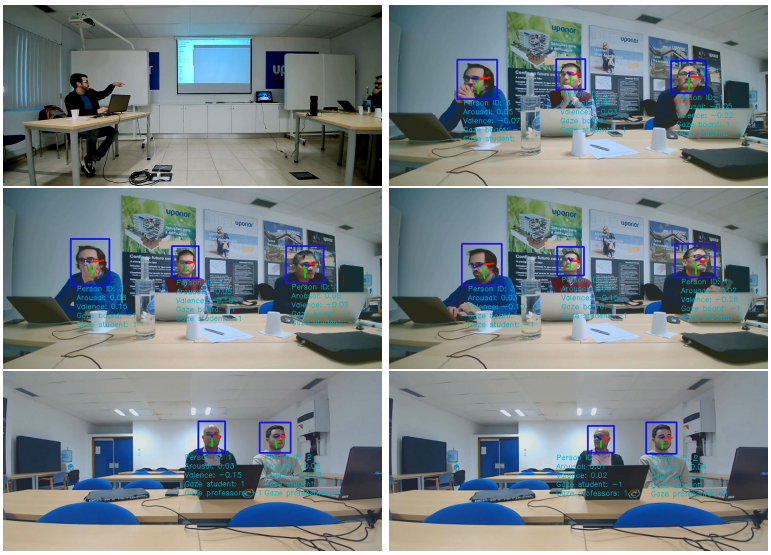

Fig. 5. Illustration of Test#2, see text.

(coordinates in meters). In the room, the student is positioned on the left and the professors, three (Group 1), on the right, being the presentation projected in the "board". Two pairs of cameras are positioned in the room. Camera_1 consists of a pair of identical cameras positioned back-to-back, with one focusing on the student and the other on the professors. Similarly, Camera_2 features the same setup, with one camera focusing on the presentation and the other on the audience, which comprises two individuals (Group 2).

Figure 5 showcases top to bottom, left to right: the general representation of the room, showing the student, the board and one of the professors; The next 3 images illustrate the professors in different situations, namely, one looking at the student and two at the board, one looking to the student and two to their computers, and one looking at the computer, one at the board and one at the student; The last two images illustrates the audience, two persons looking at the student, and one looking at the student and one at the professors.

Figure 6 (top-left) shows the valence and arousal values per frame for the individual with ID#3 (Professor_1). This person exhibits consistently low arousal,

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_9

indicating weak emotional intensity, while the valence fluctuates from negative to positive. These dynamics suggest an overall neutral emotional state, as the valence shows low absolute values despite its polarity shifts. The top-middle plot displays the valence and arousal data for the individual with ID#2 (Audience$_1$). Similar to ID#3, this person demonstrates low arousal, signifying weak emotional intensity. However, the valence remains predominantly negative, indicating an overall mild negative emotional state.

In the top-right figure is depicted the engagement between the person with ID#3 and the student (PoI), while bottom-left shows the engagement of the person with ID#3 with the board. These plots reveal that this person tends to focus more on the student than the board. The valence transition observed in the top-left plot, from negative to positive, is consistent with the engagement patterns, as the engagement also transitions from low to high values.

The bottom-middle plot highlights the engagement from person with ID#2 with the student (as PoI), and in the bottom-right the engagement with the professors. These plots indicate that this person, as ID#3, directs its attention more frequently to the student. Furthermore, the negative valence observed in the top plots aligns with the engagement trends in the bottom, as the engagement values remain consistently low. The above-mentioned plots reinforce the previously discussed engagement dynamics, as defined by the engagement formula outlined earlier. Engagement occurs only when the individual is looking at the point of interest, with the engagement level modulating in accordance with arousal intensity. This highlights the interplay between gaze, attention, and emotional engagement, underscoring the importance of arousal in driving changes in interaction focus.

Figure 7 illustrates the engagement trends over time for the three professors during a 2-minute presentation followed by a 2-minute arguing. Green dots represent a positive engagement, yellow a negative engagement, and red no engagement. The top plot shows the professors engagement with the board, while the bottom depicts their engagement with the student. The first 3,600 frames correspond to the student's presentation, while the remaining frames correspond to the arguing.

In terms of results, for the same 4-minutes period mentioned before, for the group 1 (professors) the engagement to the board (b) was $PE_b(P, \text{Group 1}) = 23\%$ of the time, i.e., 1,656 frames of engagement for a maximum of 7,200 (4 minutes $\times$ 60 seconds $\times$ 30 frames). To the student (s), the period engagement was $PE_s(P, \text{Group 1}) = 27\%$. Finally, the $E(\text{Group 1} \cup \text{Group 2}) = 41\%$, i.e., the groups were engaged 41% of the event, counting both the engagement from the student and board.

5 Conclusion

This paper presents a modular and scalable model for real-time emotion analysis and engagement detection, combining advanced deep learning models with multimodal data processing. By integrating valence-arousal prediction, head pose estimation, and individual identity tracking, the system achieves robust performance in diverse scenarios, such as educational and behavioural studies. The

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_9

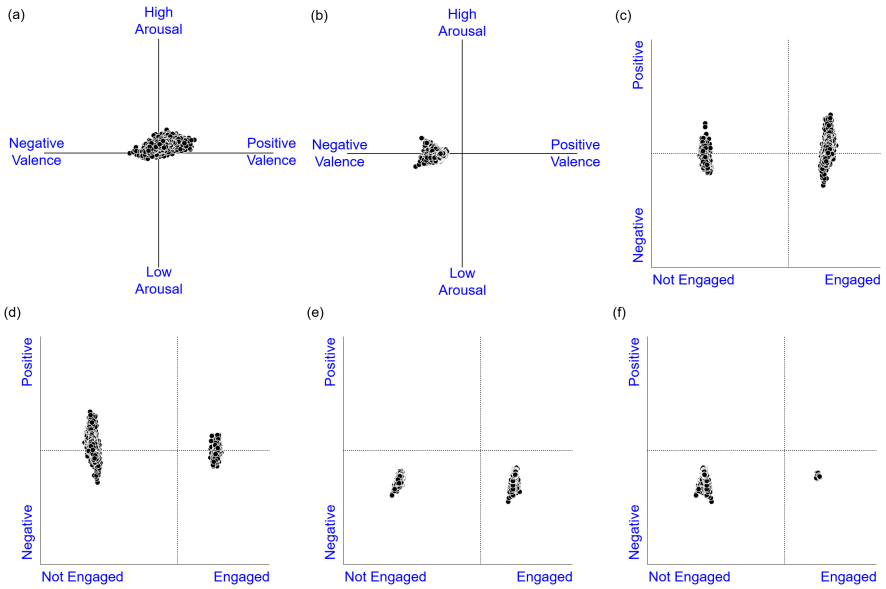

Fig. 6. (a) Frame-by-frame graph of valence and arousal for person ID#3 and (b) ID#2, engagement for person ID#3 towards the (c) student and (d) board, followed by the engagement towards the (e) student and the (f) professors for person ID#2.

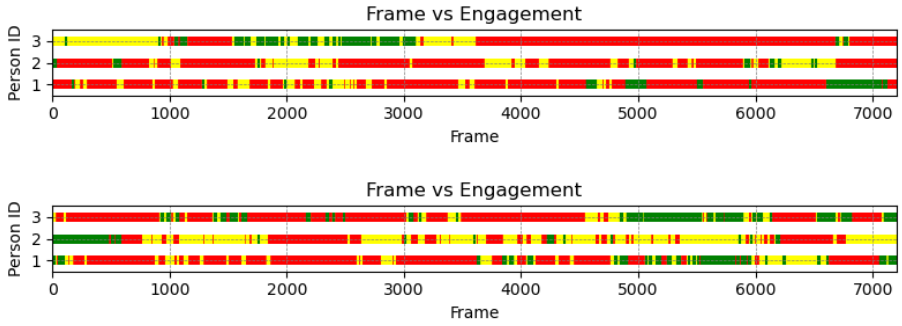

Fig. 7. Engagement of each professor, on the top, with the board, and on the bottom, with the student. See details in the text.

framework demonstrates effective classification of engagement levels, leveraging gaze, emotional states, and head orientation to provide detailed insights into individual and group behaviours.

When the camera focuses the sole PoI (e.g., Test#1), the system operates effectively even with positional changes in the room, as the camera's optical axis provides a fixed reference for yaw detection. However, challenges arise with vertical movement since pitch values change with head tilts but are not currently linked to the camera's vertical alignment. When the PoI is not the main focus of

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_9

the camera (e.g., Test#2), our model relies on fixed positions for individuals, PoI, and cameras. This limitation arises because accurate angle estimation depends on known spatial relationships. Enhancing the system to handle dynamic scenarios (e.g., moving individuals or PoI) is a future goal. This may involve integrating real-time positional tracking and incorporating changes in pitch due to vertical shifts.

Future work, in addition to the aspects already mentioned, includes integrating additional dimensions or sources such as speech and physiological data, which hopefully will further improve the model performance. However, the primary future goal is to improve adaptability to dynamic scenarios, thereby increasing the system's versatility. With these advancements, the proposed framework can become a valuable tool in fields such as human-computer interaction, offering deeper insights into human emotions and engagement.

Acknowledgments

This work is supported by UID/04516/NOVA Laboratory for Computer Science and Informatics (NOVA LINCS) with the financial support of FCT.IP, and by the project AI.EVENT: Monitor Live Audience with AI (ALGARVE-FEDER-01180500, Ref. 17325) co-financed by ALGARVE 2030, Portugal 2030 and by the European Union.

References

1. Berlincioni, L., Cultrera, L., Becattini, F., Bimbo, A.D.: Neuromorphic valence and arousal estimation. Journal of Ambient Intelligence and Humanized Computing pp. 1–11 (2024)
2. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
3. Bruin, J., Stuldreher, I.V., Perone, P., Hogenelst, K., Naber, M., Kamphuis, W., Brouwer, A.M.: Detection of arousal and valence from facial expressions and physiological responses evoked by different types of stressors. Frontiers in Neuroergonomics **5**, 1338243 (Mar 2024). https://doi.org/10.3389/fnrgo.2024.1338243
4. Dickinson, K., Caldwell, K., Graviss, E., Nguyen, D., Awad, M., Tan, S., Winer, J., Pei, K., Committee, A.E.T., et al.: Assessing learner engagement with virtual educational events: Development of the virtual in-class engagement measure (VIEM). The American Journal of Surgery **222**(6), 1044–1049 (2021)
5. Einsle, C.S., Escalera-Izquierdo, G., García-Fernández, J.: Social media hook sports events: a systematic review of engagement. Communication & Society **36**(3), 133–151 (2023)
6. Fanelli, G., Dantone, M., Gall, J., Fossati, A., Van Gool, L.: Random forests for real time 3d face analysis. International Journal of Computer Vision **101**, 437–458 (2013)
7. Gupta, S., Kumar, P., Tekchandani, R.K.: Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. Multimedia Tools and Applications **82**(8), 11365–11394 (2023)
8. Harrison, E.N.B., Kwon, W.S.: Brands talking on events? brand personification in real-time marketing tweets to drive consumer engagement. Journal of Product & Brand Management **32**(8), 1319–1337 (2023)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_9

9. Hempel, T., Abdelrahman, A.A., Al-Hamadi, A.: 6d rotation representation for unconstrained head pose estimation. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 2496–2500. IEEE (2022)
10. Hempel, T., Abdelrahman, A.A., Al-Hamadi, A.: Toward robust and unconstrained full range of rotation head pose estimation. IEEE Transactions on Image Processing **33**, 2377–2387 (2024)
11. Hwooi, S.K.W., Othmani, A., Sabri, A.Q.M.: Deep learning-based approach for continuous affect prediction from facial expression images in valence-arousal space. IEEE Access **10**, 96053–96065 (2022)
12. Lasri, I., Riadsolh, A., Elbelkacemi, M.: Facial emotion recognition of deaf and hard-of-hearing students for engagement detection using deep learning. Education and Information Technologies **28**(4), 4069–4092 (2023)
13. Mollahosseini, A., Hasani, B., Mahoor, M.H.: AffectNet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing **10**(1), 18–31 (2017)
14. Nguyen, H.H., Huynh, V.T., Kim, S.H.: An ensemble approach for facial expression analysis in video. arXiv preprint arXiv:2203.12891 (2022)
15. Rodrigues, J., Cardoso, P., Lemos, M., Cherniavska, O., Bica, P.: Engagement monotorization in crowded environments: A conceptual framework. In: 11th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion (DSAI2024). Abu Dhabi, UAE (November 2024). https://doi.org/10.1145/3696593.3696632, accepted
16. Savchenko, A.V.: Frame-level prediction of facial expressions, valence, arousal and action units for mobile devices. arXiv preprint arXiv:2203.13436 (2022)
17. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)
18. Serengil, S., Özpınar, A.: A benchmark of facial recognition pipelines and co-usability performances of modules. Bilişim Teknolojileri Dergisi **17**(2), 95–107 (2024)
19. Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., Sun, J.: CrowdHuman: A benchmark for detecting human in a crowd. arXiv e-prints pp. arXiv–1805 (2018)
20. Veltmeijer, E.A., Gerritsen, C., Hindriks, K.V.: Automatic emotion recognition for groups: a review. IEEE Transactions on Affective Computing **14**(1), 89–107 (2021)
21. Vu, T.H., Osokin, A., Laptev, I.: Context-aware CNNs for person head detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2893–2901 (2015)
22. Zhang, N., Luo, J., Gao, W.: Research on face detection technology based on MTCNN. In: 2020 international conference on computer network, electronic and automation (ICCNEA). pp. 154–158. IEEE (2020)
23. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: From facial expression recognition to interpersonal relation prediction. International Journal of Computer Vision **126**, 550–569 (2018)
24. Zhao, Z., Li, Y., Yang, J., Ma, Y.: A lightweight facial expression recognition model for automated engagement detection. Signal, Image and Video Processing **18**(4), 3553–3563 (2024)
25. Zhou, Y., Gregson, J.: WHENet: Real-time fine-grained estimation for wide range head pose. In: Proceedings of the 31st British Machine Vision Virtual Conference. pp. 1–13 (Sep 2020)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_9