RESEARCH-ARTICLE

Engagement Monitorization in Crowded Environments: A Conceptual Framework

JOÃO M.F. RODRIGUES, University of Algarve, Faro, Faro, Portugal

PEDRO J.S. CARDOSO, University of Algarve, Faro, Faro, Portugal

MARCO LEMOS, University of Algarve, Faro, Faro, Portugal

OLENA CHERNIAVSKA, Swiss Federal Institute of Technology, Zurich, Zurich, ZH, Switzerland

PAULO BICA

Open Access Support provided by:

University of Algarve

Swiss Federal Institute of Technology, Zurich

Published: 31 July 2025

DSAI 2024: 11th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion
November 13 - 15, 2024
Abu Dhabi, United Arab Emirates

Engagement Monitorization in Crowded Environments: A Conceptual Framework

João M.F. Rodrigues
NOVA LINCS & ISE, Universidade do Algarve
Portugal
jrodrig@ualg.pt

Pedro J.S. Cardoso
NOVA LINCS & ISE, Universidade do Algarve
Portugal
pcardoso@ualg.pt

Marco Lemos
NOVA LINCS & ISE, Universidade do Algarve
Portugal
a72178@ualg.pt

Olena Cherniavska
Swiss Federal Institute of Technology
ETH Zurich
Switzerland
helenacherniavska@gmail.com

Paulo Bica
SPIC
Portugal
pbica@spic.pt

Abstract

Accessibility has emerged as a fundamental aspect of software development, aiming to ensure that digital experiences are inclusive and usable by individuals of all abilities. Humans are prepared to comprehend others' emotional expressions from subtle body movements or facial expressions. Additionally, emotions and sentiments lie on the basis of group behaviours, influencing how we interact, cooperate, and form social bonds within communities. Detecting audience engagement in events in real-time means monitoring emotions, sentiments, behaviours, attention, and scene dynamics in group(s) and crowds. In this context, this position paper introduces a conceptual framework for engagement monitorisation in crowded environments, outlining the methodology, metrics, and architecture to do this monitorisation.

CCS Concepts

• **Human-centered computing**; • **Applied computing**; • **Computing methodologies**;

Keywords

Additional Keywords and PhrasesEngagement, Emotion, Behaviours, Scene Dynamics, Crowd, Metrics

ACM Reference Format:

João M.F. Rodrigues, Pedro J.S. Cardoso, Marco Lemos, Olena Cherniavska, and Paulo Bica. 2024. Engagement Monitorization in Crowded Environments: A Conceptual Framework. In *11th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion (DSAI 2024), November 13–15, 2024, Abu Dhabi, United Arab Emirates*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3696593.3696632

DSAI 2024, November 13–15, 2024, Abu Dhabi, United Arab Emirates

ACM ISBN 979-8-4007-0729-2/24/11
https://doi.org/10.1145/3696593.3696632

1 Introduction

In addition to solutions designed specifically for people with disabilities, the disability technological sector offers goods and services that improve accessibility for all users, whether impaired or not. Digital technologies, such as big data analysis and artificial intelligence (AI) can help to improve the user experience in events and return feedback to the promoters, allowing the latter to fine-tune the event to meet attendees' expectations. This improves resource management and enhances personalized experiences for events and other activities.

In the same context, human-machine collaboration (HMC) requires that machines in the broadest sense are designed to work together or learn how to work with humans. This means that hardware, software, and interfaces must be able to assess the users' unique needs and behaviours, as well as the context in which they are used and the surrounding environment, to enable on-the-spot cooperation with humans.

Although there are several definitions for "engagement", here the focus is on real-world event engagement (not online or virtual). This case is also known as audience engagement, which refers to an event's ability to hold the attention of its attendees as well as promote their participation. Real-world engaging events are expected to be captivating and compelling; that is, they are expected to capture the attention of their audience from the start and maintain it consistently, or at least during specific key points or periods of the event.

Although several studies emphasise personal engagement or group engagement online (such as social media engagement [8], consumer engagement using tweets [10], learner engagement with virtual educational events [7], attendee engagement, and emotional attachment in cultural events [17]), to the best of our knowledge, none focus on real-world/real-time engagement in (indoor and outdoor) live events.

In this context, constructing a framework for detecting audience engagement in events is deeply related to HMC, and means, not only monitoring emotions and sentiments in groups, as emotions lie at the basis of the behaviour of a group [37], but also includes analysing data on the organic movement of the group(s) in the

real-world environment, i.e., observing how groups of people move through a crowd and spot trends or variations [35].

At this point, it is important to define two concepts already mentioned in the text, *crowd* and *group* [30]. A (i) *group* consists of a collection of people that can range from a size of two persons to hundreds, in mutual presence at a given moment, who are having some form of social interaction. Its members are close to each other, with a similar speed and with a similar direction of motion. In an event, several groups can coexist. On the other hand, (ii) *crowd (or mass)* is a unique large group of individuals sharing a common physical location. It is usually formed when people with the same goal become one single entity, losing their individuality and adopting the behaviour of the crowd entity.

It is important to stress that here, we are not going to analyse or discuss psychological theories of crowd behaviour. For that, e.g., we refer to the work of Varghese and Thampi [36]. Furthermore, unless otherwise noted, the word "group" will be used generically to refer to groups and crowds in the remaining text.

Developing a framework for detecting audience engagement in live events in real-time, includes analysing emotion, sentiment, behaviour, and attention mechanisms [23] within crowds, in individual groups, or the collection of groups in the event. That framework should also include assessing scene dynamics, such as creating crowd density maps and crowd-counting [5].

The paper's main contribution is the definition of the initial, methodology, metrics, and architecture for crowd/group engagement monitorisation in real-world environments/events. After the initial introduction (this section), section 2 presents a brief contextualisation and state of the art, section 3 the proposed crowd engagement detector prototype, and section 4 the conclusions and future work.

2 Contextualization and State-of-the-Art

The field of affective computing (AffC) focuses on integrating the understanding of emotions, affects and sentiments [36]. Recent years have seen significant advancements in the discipline, although they have mostly concentrated on stand-alone persons. The most researched human emotional recognition pathways are speech, body language, facial expressions, and physiological indicators. Nevertheless, some of these methods do not apply to crowds and groups, or present huge limitations [2].

Generically, there are two primary categories of methods for approaching emotion and behaviour analysis and sequentially engagement analysis in groups and/or crowds: (i) *Microscopical* (or bottom-up) methods, where people in the crowd are regarded as a group of individuals [37], i.e., individuals in the "video" are examined, and the information gleaned from these studies is then utilized to extrapolate information at the collective level; (ii) *Macroscopical* (or top-down) methods, composed by holistic procedures that regard the crowd as one cohesive entity, rather than needing to track and segregate each person separately [37].

Microscopic methods often work best in scenarios where small groups can be reliably tracked, that is, when there are few occlusions, low density, and clear visibility of people. When population density rises, and tracking quality sharply declines, macroscopic techniques are more appropriate. In this paper, we will focus on an architecture that integrates both methods, individually and as complements of each other.

The study by Rabiee *et al.* [25] is to the best of our knowledge the first work tackling the problem of emotion detection in large crowds. Their approach followed for classification is macroscopic, i.e., based on scene features and not on individuals, and considers scene dynamics. Zhang *et al.* [39] work with video data for groups, being emotion motion features extracted from coherent motion patterns, for both arousal and valence separately, which are then mapped to the arousal-valence plane. Other solutions point to hybrid methods by combining approaches that can be completed by focusing on different aspects of an image, for example, by employing information from faces and entire scenes, or also including body information. Most studies combine face-level analysis with scene-level analysis. For instance, Huang *et al.* [11] combine faces, scenes, and upper bodies; Nagarajan *et al.* [20] combine faces, scenes, and places; and Li *et al.* [15] combine faces, scenes, bodies, and skeletons. Rathod *et al.* [26] present a survey on perceived group sentiment analysis. For crowd density maps, and a review of existing methods see, e.g., the works of Patwal *et al.* [24] or Khan *et al.* [12]. For human behaviour, including abnormal actions, group activities, and other types of human activities, we refer to, e.g., the studies of Surek *et al.* [33] or Morshed *et al.* [18].

In the case of crowd metrics, the number of publications is not significant. For example, Sánchez *et al.* [30] present metrics for crowd behaviour, while Bendali-Braham *et al.* [3] present metrics and solutions for crowd analysis to monitor crowd events, but with the focus of surveillance (which is not the focus of this paper). Wirth *et al.* [38] present an interesting discussion about the neighbourhood of interaction in human crowds. Khan *et al.* [13] present a paper about visual crowd analysis. They address the open research challenges in this field and divide the field into six main categories: crowd counting, object detection and tracking, motion analysis, behaviour recognition, anomaly detection, and crowd prediction. For these six categories, they concentrate (conduct a state-of-the-art) on the most recent deep learning techniques and, based on the category, they provide a presentation of the current model architectures, training strategies, evaluation metrics, loss functions etc.

In summary, the literature on group and crowd engagement is still narrow. Nevertheless, there is more than enough evidence in the literature that there are paths to be followed that can result in an engagement monitorization architecture.

Also important, is to validate the existence of datasets for the development of this research. In this context, some datasets exist that fit different behaviour purposes, such as (but are not limited to): the UCSD Pedestrian dataset[1], the CUHK Avenue dataset[2], the UMN dataset[3], the ShanghaiTech Campus dataset[4], the BEHAVE

[1]UCSD dataset available at (accessed 2024/05/07): http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm
[2]CUHK Avenue dataset available at (accessed 2024/05/07): http://www.cse.cuhk.edu.hk/leojia/projects/detectabnormal/dataset.html
[3]UMN dataset available at (accessed 2024/05/07): http://mha.cs.umn.edu/proj_events.shtml#crowd
[4]ShanghaiTech Campus dataset available at (accessed 2024/05/07): https://svip-lab.github.io/dataset/campus_dataset.html

dataset[5], the BOSS dataset[6], the UT Interactions dataset[7], and the UCF-Crime dataset[8].

One of the main bottlenecks to advancing the field/project is the scarcity of public datasets showing videos involving groups or crowds and providing emotional or sentiment annotations. Nevertheless, they exist, such as the cases of HAPPEI[9] - (HAPpy People Images dataset) and MED[10] (Motion Emotion Dataset). Other datasets exist, like GAFF 2.0 and 3.0 [6], which are annotated for the emotional valence, MultiEmoVA [19], annotated for classes arousal (low, medium, and high) and valence (negative, neutral, and positive), Emotic [14] which has 26 emotion labels plus arousal valence and dominance, and GroupEmoW [9] with the class valence. For a more detailed description of these datasets see the work of Veltmeijer *et al.* [37]. More crowd datasets (11) can be found in Varghese and Thampi work [36].

For counting the number of people in a crowd, several datasets are available, e.g., see the work of Patwal *et al.* [24]. For a survey about emotions and sentiment datasets for single persons or small groups (less than 4 persons) please see [16], a publication from the authors of the present paper. In this case, more than 100 datasets were identified, around 70% of them public datasets. The next section presents the methodology and the architecture for monitoring engagement for crowds.

3 Engagement Monitorization in Crowded Environments

This section is divided into three subsections. The first focuses on the (i) *methodology* for event engagement analysis, the second proposes the (ii) *metrics* for engagement monitoring, and, finally, an initial (iii) *architecture* to acquire information and quantify the engagement level is presented.

3.1 Methodology

Whether a macroscopic or microscopic technique is used, the following four phases make up the pipeline for generic group/crowd engagement analysis:

(i) *Detection.* Its goal is to pinpoint the locations of groups of people (microscopic environments) and the mass of people / crowds (macroscopic strategies) in every frame. This step has received a lot of research attention, and there are already some highly accurate and performant detection models available [2].

(ii) *Feature extraction.* It extracts information and calculates a range of metrics that characterise the emotional state, topology, and scene dynamics. These metrics may be tracked throughout time and calculated at two different levels: the (ii.1) individual level, microscopic, which considers each person as an independent entity that contributes to the group; and the (ii.2) crowd level, macroscopic, which considers each mass of people as a single entity. Valence, arousal monitoring, crowd density, and motion patterns are a few examples of features to be extracted (but not the only ones; see next sub-section).

(iii) *Monitoring.* It attempts to detect changes along the time in individual group(s) and crowds, in a series of successive frames, pinpointing specific alterations (emotional, behavioural etc.), and a timeline with the engagement along the duration of the event. Additionally, the groups' (plural) and the crowd (single) dominant "flows" are established, and the scene dynamics are monitored.

(iv) *Classification.* This final step combines the attributes extracted in (ii) and the monitorization in the previous step, (iii), to identify engagement levels in "video" sequences. The engagement level will be divided by *instantaneous*, (pre-)defined *period*(s), and *event*. In more details, instantaneous engagement is the engagement level at a specific time, period engagement is the engagement level during a specific period (window) of time, and event engagement is the engagement level during the entire event.

(v) *Presentation and reporting.* The final step is to present the results in a user-friendly way (e.g., a dashboard) and to report the results to the event promoters. The results can be presented in real-time or after the event. As mentioned, real-time results can be used to adjust the event on-the-fly, while post-event results can be used to improve future events.

The engagement research strategy broadly follows the 5 phases presented above. Nevertheless, two channels of research should occur in parallel: *Microscopic channel*, more focused on groups, where the combination of individual attributes of each element of the group and the environment will be used to characterize each group. *Macroscopic channel*, which is more focused on crowds, in this case, the whole image will be the main input, as well as the surrounding environment. Finally, these two channels will be combined for the final engagement classification.

3.2 Metrics

The metrics could be divided into five main groups: (a) emotion, (b) sentiment, (c) behaviour, (d) attention, and (e) scene dynamics. Each group has to be divided into microscopic (group) and macroscopic (crowd) levels. But let us give a brief explanation of each group and its subgroups.

(a) Emotion (E) [29] – to compute the emotion level it is considered: ***valence*** (EV – pleasantness, affective state, affect appraisals – goes from unpleasant/negative to pleasant/positive), ***arousal*** (EA – activation, energy and stimulation level, physiological reactions – goes from deactivation to activation), and ***valence-arousal plain*** (EP). The plane is divided into 4 quadrants, with the first related to "joy/happiness" (e.g., pleased, happy, or excited); the second associated with "anger/fear" (e.g., annoying, angry, or nervous); the third associated with "sadness" (e.g., sad, bored, or sleepy); and the fourth related with "pleasure/tenderness" (e.g., calm, peaceful, or relaxed). In the context of the previous idea, this metric is still divided into:

(a.1) Microscopic – *group(s)* (g) – retrieved from the combination of *individual* (person) information within the group: EVg_n,

[5]BEHAVE dataset available at (accessed 2024/05/07): http://groups.inf.ed.ac.uk/vision/BEHAVEDATA/INTERACTIONS/

[6]BOSS dataset available at (accessed 2024/05/07): http://velastin.dynu.com/video-datasets/BOSSdata/index.html

[7]UT Interactions dataset available at (accessed 2024/05/07): https://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html

[8]UCF-Crime dataset available at (accessed 2024/05/07): https://web-pages.uncc.edu/cchen62/dataset.html

[9]HAPpy People Images dataset available at (accessed 2024/05/07): https://cs.anu.edu.au/few/Group.htm

[10]MED Motion Emotion Dataset available at (accessed 2024/05/07): https://github.com/hosseinm/med?tab=readme-ov-file

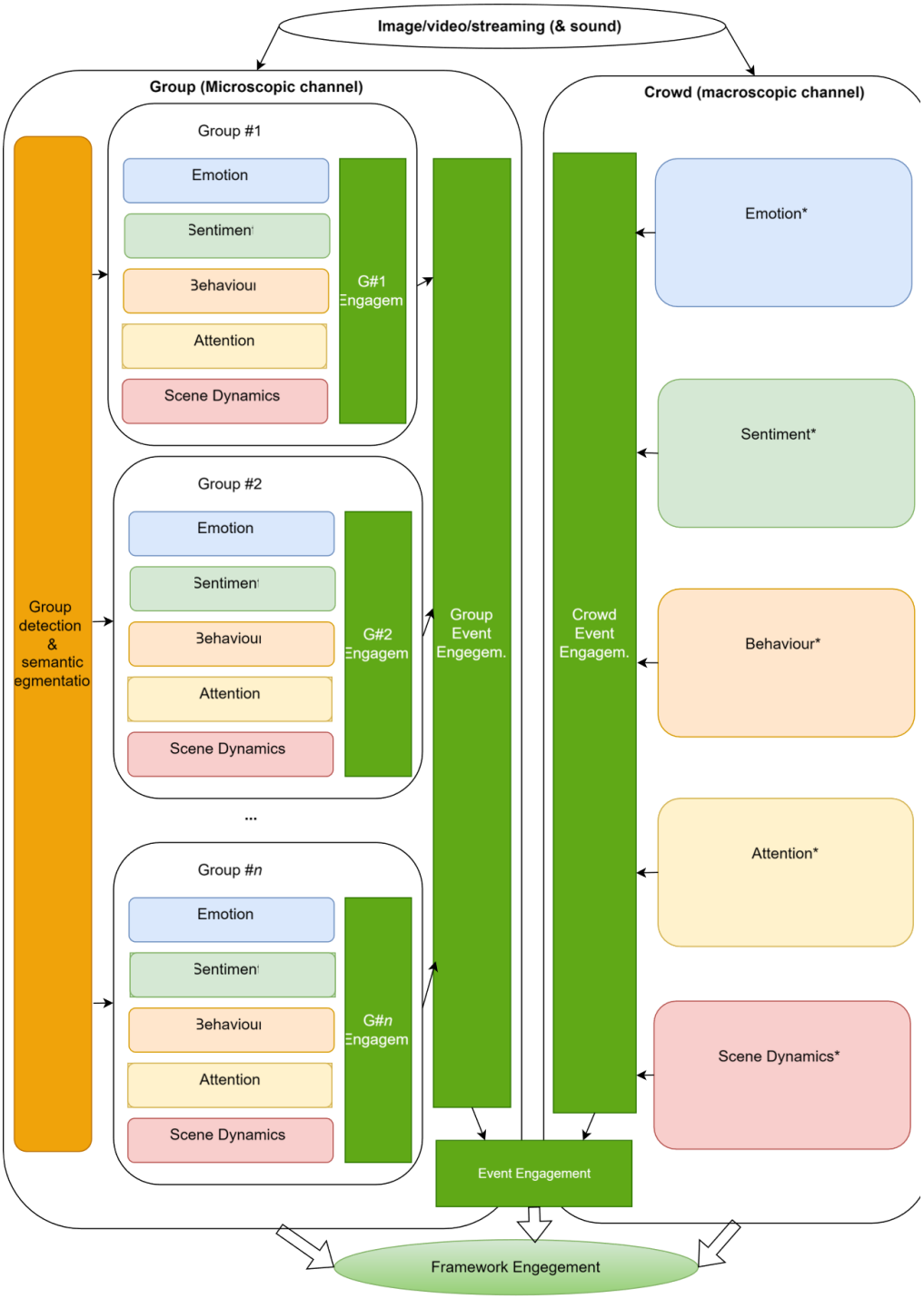

Figure 1: - Block diagram of the engagement monitorization.

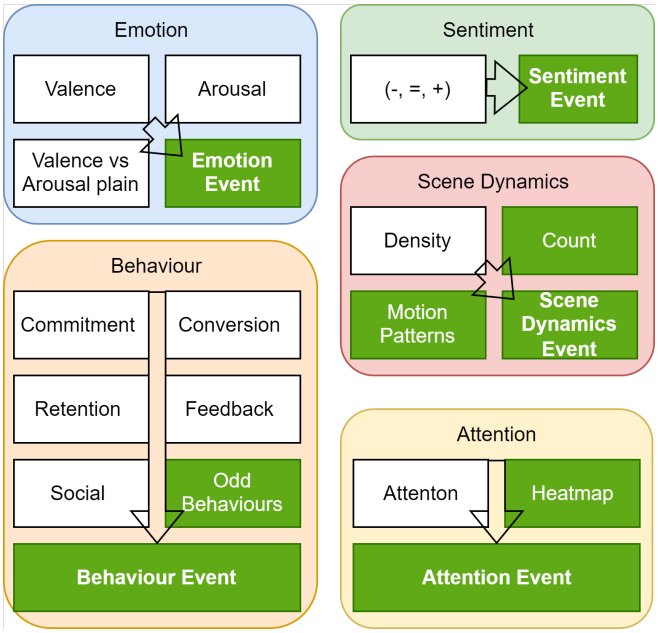

Figure 2: - Sub-block diagram of the engagement monitorization framework.

EAg_n, and EPg_n, where n is the group's identification number (from different groups inside the event);

(a.2) Macroscopic – crowd (c) – retrieving: EVc, EAc, and EPc;

(a.3) Event (e) – retrieved from the combination of *group(s)* and *crowd* information within the event: $EVe = C\{\oplus_n EVg_n, EVc\}$, $EAe = C\{EAg_n, EAc\}$,and $Epe = C\{EPg_n, EPc\}$, where $\oplus_n$ is the combination of the information retrieved from the different groups and C is the combined information from different variables. Being the emotion engagement for the event $EVe = C\{EVe, EAe, EPe\}$.

(b) Sentiment (S) [1] – For sentiment three levels are considered (namely, negative, neutral, and positive), which are to be divided in the same paths as emotions, i.e.:

(b.1) Microscopic – Sg$_n$;

***(b.2)* Macroscopic** – Sc;

(b.3) Event – $Se = \mathbb{C}\{\oplus_n Sg_n, Sc\}$

(c) Behaviour (B) – Behaviour examples include, but are not limited to, the following categories, which must be divided into *microscopic* (g) and *macroscopic* (c):

(c.1) Commitment (ct) – track how many times a group or the crowd interacts with a product, service, or event – Bct, which is computed as the ratio between the number of times and the duration of the event, resulting in the indicator $\{Bct = \mathbb{C}\{\oplus_n Bctg_n, Bctc\}$.

(c.2) Conversion (cv) – track how many groups or if the crowd completes a pre-defined action/path – Bcv, computed as the ratio between the number of paths repeated and the duration of the event, returning the indicator $Bcv = \mathbb{C}\{\oplus_n Bcvg_n, Bcvc\}$.

(c.3) Retention (rt) – track how many groups or if the crowd return to a product, service, or event sector after their first visit – Brt, computed as the ratio between the number of returns and the duration of the event, $Brt = \mathbb{C}\{\oplus_n Brtg_n, Brtc\}$.

(c.4) Feedback (fb) – track how many times a type (t) of feedback was presented by the group or crowd (e.g., waving or clapping hands) – Bfb_t, the ratio between the number of times *per* type and the duration of the event, $Bfb_t = \mathbb{C}\{\oplus_n Bfb_{t,g_n}, Bfb_{t,c}\}$.

(c.5) Social (s) – track how many different groups interact with each other – B$s_{i,j}$, the ratio between the number of interactions and the duration of the event between group i and j, $Bs = \mathbb{C}\{\oplus_{i,j} Bs, i, j\}$.

(c.6) Odd Behaviours / alerts (ob) – detect a subset or group of attendees with a behaviour very different from the other groups or event sectors or abnormal crowd behaviour – Bob. E.g., a group running in a specific direction or to a specific sector of the event, or one group clapping hands while the remaining groups do not. Being the odd behaviours of the event $Bob = \mathbb{C}\{\oplus_n Bobg_n, Bobc\}$.

***(c.7)* Event** – the event behaviour is then computed as $Be = \mathbb{C}\{Bct, Bcv, Brt, Bfb_t, Bs, Bob\}$.

(d) Attention (At) - to compute the attention level, are considered, but not limited to, the following categories, which must be divided into *microscopic* and *macroscopic*:

(d.1) Focus, track in the group/crowd if the head direction (gaze) is pointing to the event's main focus. It represents the ratio between persons/groups paying attention and all the persons/groups present, $AtF = \mathbb{C}\{AtFg_n, AtFc\}$.

(d.2) Heatmaps, present the maps of the sum of At along the event, $Athm = \mathbb{C}\{Athmg_n, Athmc\}$.

***(d.3)* Event**, $Ate = \mathbb{C}\{AtF, Athm\}$

(e) Scene dynamics (SD) – computation includes, but is not limited to (and again they have to be divided into microscopic and macroscopic):

(e.1) Density, represents the density of persons by section in the event (SDd). It also creates a density map, with the sum of SDd along the time of the event (SDdm); $SDd = \mathbb{C}\{\oplus_n SDdg_n, SDdc\}\}$ and $SDdm = \mathbb{C}\{\oplus_n SDdmg_n, SDdmc\}$.

(e.2) Count, track during the event the number of persons assisting the event, SDk.

(e.3) Motion patterns, tracks during the event the different motion patterns (SDm). It also creates a heatmap with the motion patterns of the groups and respective chronology of how groups move inside the event (SDhm), returning $SDm = \mathbb{C}\{\oplus_n SDmg_n, SDmc\}$ and $SDhm = \mathbb{C}\{\oplus_n SDhmg_n, SDhmc\}$.

(e.4) Event, $SDe = \mathbb{C}\{SDd, SDdm, SDk, SDm, SDhm\}$.

The *instantaneous engagement* in the event at any time t is given by $E_t = \mathbb{C}\{Ee, Se, Be, Ate, SDe\}$, the *period engagement* by $E_p = \oplus_{t\in[ti,tf]} E_t$ (where ti and $tf, ti < tf,$ are to different times in the event timeline), and the *event engagement* $E = \oplus_{t\in I} E_t$, i.e., it accounts for the entire event, with duration interval I.

The next section presents an explain the generic architecture of the crowded engagement monitorization framework.

3.3 Architecture

Figure 1 presents the global block diagram of the engagement framework. On the left are the blocks related to the microscopic channel (group engagement), and on the right the macroscopic channel (crowd engagement). Since they have the same designation, despite being completely different, the macroscopic are marked with a "*" symbol.

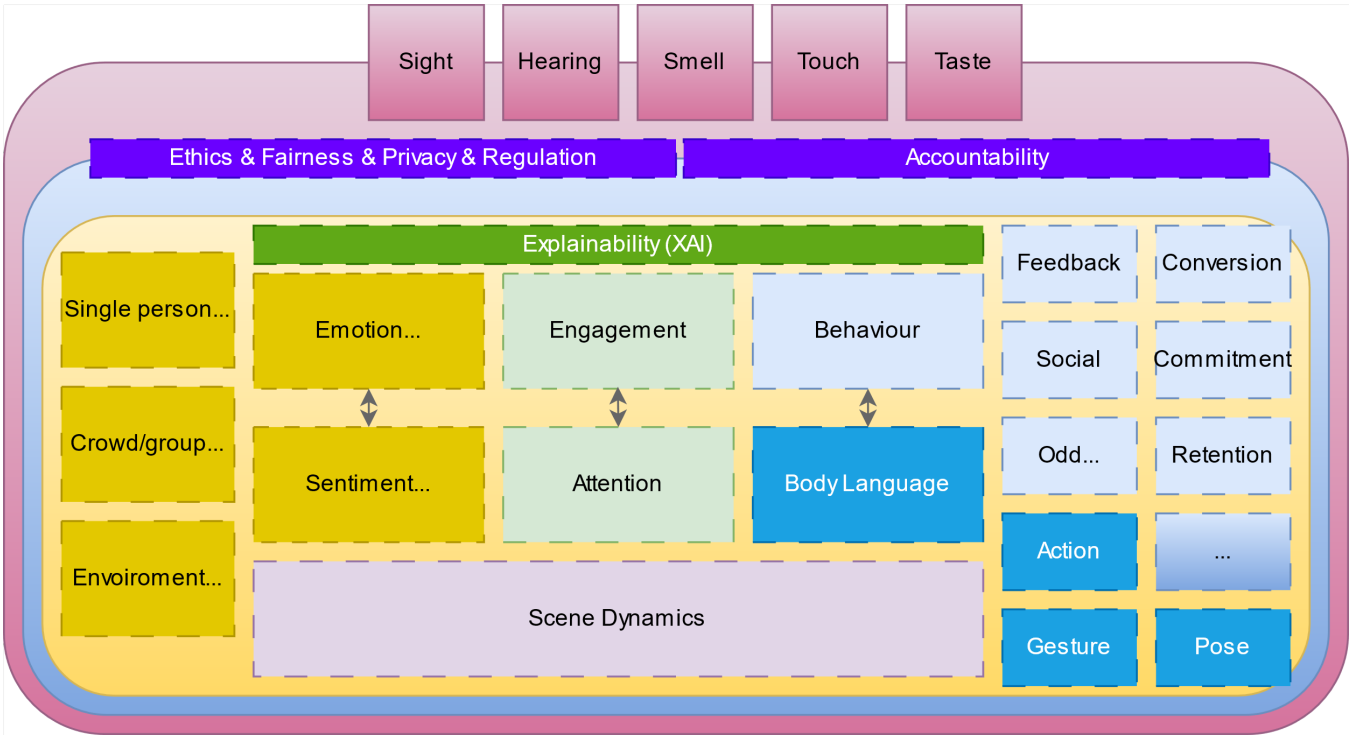

Figure 3: - Overall architecture of CAffC, a unified emotion-sentiment-engagement-behavioural framework.

Going back to the left block, image/video/streaming are initially processed to detect where are the groups. For that purpose, the advised solution is the use of semantic segmentation.

Having the groups segmented and located in the event's space, for each group 5 main modules a processed in parallel, namely: (a) emotion, (b) sentiment, (c) behaviour, (d) attention, and (e) scene dynamics. Different sub-modules are computed for each of modules (a) to (e) to generate the respective metrics. Then, the individual metrics for each group are combined to return the metrics of the event looking only at the group aspect (Figure 2).

In parallel with the group processing, the entire image/video/streaming is analysed to detect the crowd and the same 5 main modules are processed concurrently. It is important to stress that despite the metrics being the same, the sub-modules to process/achieve those metrics are completely different. In the case of microscopic information we are dealing with the information retrieved at the level of the individual/group, which is then combined. In the other case, macroscopic case, we are dealing with the global information retrieved from the image/video/streaming.

Finally, information from the microscopic channel (groups) and the macroscopic channel (crowd) are combined to return the *event engagement* block.

It is important to notice that to characterise the *engagement framework*, not only the final engagement is important, i.e., the engagement of each group and some key metrics are fundamental, all of which are marked in green in Figure 1 and Figure 2. In this sense, for real-time systems, the architecture should be able to process the information live and return the engagement level in real-time. This means that the system should be able to process the information in a few milliseconds and return the engagement level in a few seconds. This is a challenge, but it is achievable with the current state of the art in AI and computer vision. Furthermore, monitoring the engagement level in real-time can be used to adjust the event on-the-fly, for example, by changing the music, the lights, or the speaker, to increase the engagement level. This can be done automatically by the system, or manually by the event promoter, who can use the system to get recommendations on how to improve the event.

Information must therefore be provided in a meaningful way so that the event promoter can understand it and act on it. This can be done by presenting the information in a dashboard, with graphs and charts that show the engagement level over time, and the factors that influence it.

Finally, the system can also be used to improve future events, by analysing the engagement level of past events, and identifying what worked and what didn't work. This can help the event promoter to fine-tune the event, and make it more engaging for the attendees.

4 Conclusion and Future Work

This position paper presented a conceptual framework for engagement monitorisation in crowded environments. The flow of the methodology, metrics, and architecture was presented.

The methodology is divided into five main phases: detection, feature extraction, monitoring, classification, and presentation and reporting. Further, the methodology is divided into two channels, microscopic (group) and macroscopic (crowd), i.e., the information

is processed at the individual level and the global level. Immediate challenges at the individual level are related to the occlusions, low density, and clear visibility of people, while at the global level are related to population density and tracking quality.

The metrics are divided into five main groups: emotion, sentiment, behaviour, attention, and scene dynamics. These groups are also, in general, divided into microscopic (group) and macroscopic (crowd) levels. The metrics are used to compute the engagement at different time levels, such as instantaneous, selected period, and event engagement.

Finally, the architecture is divided into two main channels, microscopic (group) and macroscopic (crowd), and each channel has the same five main modules, namely: emotion, sentiment, behaviour, attention, and scene dynamics computation. The information from the two channels is combined to return the engagement level of the event. The architecture is designed to process the information in real-time and return the engagement level live, allowing the event promoter to adjust the event on-the-fly, and to improve future events.

In summary, this position paper is not focused on the computational development of each block and sub-block, i.e., it focuses on a conceptual framework for engagement monitorisation in crowded environments, presenting the methodology, metrics, and architecture to do this monitorisation.

Future work consists of developing each of those blocks and submodules. Nevertheless, some initial (partial) work already started, namely in [4][16][21][22][28][31][32][34].
In reality, despite this paper's focus only on the engagement model, a depth architecture is being implemented, see Figure 3. It focuses on 3 main levels: (i) first level, the interaction with the five human senses, e.g. [27], where the five human senses (sight, hearing, smell, touch, and taste) are the base to extract the inputs or to return the outputs of the model. (ii) A second level is the analysis and validation of the ethics, fairness, privacy, regulation, and accountability of the global framework and each (individual) model. (iii) In the last level (third level), is the development of the different models and the direct connections and indirect connections between them to develop a unified emotion-sentiment-engagement-behavioural framework, i.e., a ***Comprehensive Affective Computing*** (CAffC) framework.

Acknowledgments

This work is supported by NOVA LINCS ref. UIDB/04516/2020 (https://doi.org/10.54499/UIDB/04516/2020) and ref. UIDP/04516/2020 (https://doi.org/10.54499/UIDP/04516/2020) with the financial support of FCT.IP, and project AI.EVENT: Monitor Live Audience with AI.

References

[1] Alslaity, A., and Orji, R. 2024. Machine learning techniques for emotion detection and sentiment analysis: current state, challenges, and future directions. Behaviour & Information Technology, 43, 1, 139-164.

[2] Bellomo, N., Liao, J., Quaini, A., Russo, L., and Siettos, C. 2023. Human behavioral crowds: Review, critical analysis, and research perspectives. Mathematical Models and Methods in Applied Sciences 33, 8, 1611–1659.

[3] Bendali-Braham, M., Weber, J., Forestier, G., Idoumghar, L., and Muller, P. A. 2021. Recent trends in crowd analysis: A review. Machine Learning with Applications 4, 100023.

[4] Cardoso, P.J.S., Rodrigues, J.M.F., and Novais, R. 2023. Multimodal Emotion Classification Supported in the Aggregation of Pre-trained Classification Models. In: Mikyška, J., de Mulatier, C., Paszynski, M., Krzhizhanovskaya, V.V., Dongarra, J.J., Sloot, P.M. (eds) Computational Science – ICCS 2023. Lecture Notes in Computer Science, vol 10477. Springer, Cham. DOI: 10.1007/978-3-031-36030-5_35

[5] Deng, L., Zhou, Q., Wang, S., Górriz, J. M., and Zhang, Y. 2023. Deep learning in crowd counting: A survey. CAAI Transactions on Intelligence Technology. https://doi.org/10.1049/cit2.12241

[6] Dhall, A., Joshi, J., Sikka, K., Goecke, R., and Sebe, N. 2015. The more the merrier: Analysing the affect of a group of people in images. In 2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG), Vol. 1, pp. 1-8. IEEE.

[7] Dickinson, K. J., Caldwell, K. E., Graviss, E. A., Nguyen, D. T., Awad, M. M., Tan, S., ... and ASE Educational Technology Committee. 2021. 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.

[8] Einsle, C. S., Escalera-Izquierdo, G., and García-Fernández, J. 2023. Social media hook sports events: a systematic review of engagement. Communication & Society, 36, 3, 133-151.

[9] Guo, X., Polania, L., Zhu, B., Boncelet, C., and Barner, K. 2020. Graph neural networks for image understanding based on multiple cues: Group emotion recognition and event recognition as use cases. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2921-2930.

[10] Harrison, E. N. B., and Kwon, W. S. 2023. Brands talking on events? Brand personification in real-time marketing tweets to drive consumer engagement. Journal of Product & Brand Management, 32, 8, 1319-1337.

[11] Huang, X., Dhall, A., Goecke, R., Pietikäinen, M., and Zhao, G. 2018. Multimodal framework for analyzing the affect of a group of people. IEEE Transactions on Multimedia 20, 10, 2706–2721.

[12] Khan, M. A., Menouar, H., and Hamila, R. 2023a. Revisiting crowd counting: State-of-the-art, trends, and future perspectives. Image and Vision Computing 129, 104597.

[13] Khan, M. A., Menouar, H., and Hamila, R. 2023b. Visual crowd analysis: Open research problems. AI Magazine 44, 3, 296–311.

[14] Kosti, R., Alvarez, J. M., Recasens, A., and Lapedriza, A. 2017. Emotion recognition in context. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1667-1675.

[15] Li, D., Luo, R., and Sun, S. 2020. Group-level emotion recognition based on faces, scenes, skeletons features. In Eleventh International Conference on Graphics and Image Processing (ICGIP 2019), SPIE, 46–51.

[16] Martins, P.V., Cardoso. P. J.S., and Rodrigues, J.M.F. 2024. Affective Computing Databases: In-depth Analysis, submitted to IEEE Transactions on Affective Computing.

[17] Meeprom, S., and Fakfare, P. 2021. Unpacking the role of self-congruence, attendee engagement and emotional attachment in cultural events. International Journal of Event and Festival Management, 12, 4, 399-417.

[18] Morshed, M. G., Sultana, T., Alam, A., and Lee, Y. K. 2023. Human action recognition: A taxonomy-based survey, updates, and opportunities. Sensors 23, 4, 2182.

[19] Mou, W., Celiktutan, O., and Gunes, H. 2015. Group-level arousal and valence recognition in static images: Face, body and context. In 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Vol. 5, pp. 1-6. IEEE.

[20] Nagarajan, B., and Oruganti, V. R. M. 2019. Group emotion recognition in adverse face detection. In 14th IEEE International Conference on Automatic Face & Gesture Recognition, IEEE, 1–5.

[21] Novais, R., Cardoso, P.J.S., and Rodrigues, J.M.F. 2023. Facial Emotions Classification Supported in an Ensemble Strategy. In: Antona, M., Stephanidis, C. (eds) Universal Access in Human-Computer Interaction. Novel Design Approaches and Technologies. HCII 2022. Lecture Notes in Computer Science, vol 13308, pp.477-488. Springer, Cham. DOI: 10.1007/978-3-031-05028-2_32

[22] Novais, Rui, Cardoso, Pedro J.S. and Rodrigues, João M. F. 2023. Emotion Classification from Speech by an Ensemble Strategy.In Procs of the 10th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion (DSAI '22). Association for Computing Machinery, New York, NY, USA, 85–90. DOI: 10.1145/3563137.3563170

[23] Novin, S., Fallah, A., Rashidi, S., and Daliri, M. R. 2023. An improved saliency model of visual attention dependent on image content. Frontiers in Human Neuroscience 16, 862588.

[24] Patwal, A., Diwakar, M., Tripathi, V., and Singh, P. 2023. Crowd counting analysis using deep learning: A critical review. Procedia Computer Science 218, 2448–2458.

[25] Rabiee, H., Haddadnia, J., Mousavi, H., Nabi, M., Murino, V., and Sebe, N. 2016. Emotion-based crowd representation for abnormality detection. arXiv preprint arXiv:1607.07646.

[26] Rathod, B., Vanzara, R., and Pandya, D. 2023. A recent survey on perceived group sentiment analysis. Journal of Visual Communication and Image Representation, 103988.

[27] Rodrigues, J. M. F., Ramos, C., Pereira, J., Cardo, J., and Cardoso, P. J. S. 2019 Five Senses Augmented Reality System: Technology Acceptance Study, IEEE Access, vol. 7, pp. 163022-163033. DOI: 10.1109/ACCESS.2019.2953003

[28] Rodrigues, J.M.F. and Cardoso, P.J.S. 2023. Body-Focused Expression Analysis: A Conceptual Framework. In: Antona, M., Stephanidis, C. (eds) Universal Access in Human-Computer Interaction. HCII 2023. Lecture Notes in Computer Science, vol 14021. Springer, Cham. DOI: 10.1007/978-3-031-35897-5_42

[29] Russell, J. A. 1980. A circumplex model of affect. Journal of personality and social psychology, 39, 6, 1161.

[30] Sánchez, F. L., Hupont, I., Tabik, S., and Herrera, F. 2020. Revisiting crowd behaviour analysis through deep learning: Taxonomy, anomaly detection, crowd emotions, datasets, opportunities and prospects. Information Fusion 64, 318–335.

[31] Santos J., Martins I., Rodrigues J.M.F. 2021. Framework for Controlling KNX Devices Based on Gestures. In: Antona M., Stephanidis C. (eds) Universal Access in Human-Computer Interaction. Access to Media, Learning and Assistive Environments. HCII 2021. LNCS 12769. Springer, Cham. DOI: 10.1007/978-3-030-78095-1_37

[32] Silva, N., Cardoso, Pedro J.S., and Rodrigues, João M.F. 2024. Sentiment Classification Model for Landscapes, Accepted to 18th International Conference on Universal Access in Human-Computer Interaction, part of HCI International 2024, 29 June - 4 July 2024, Washington DC, USA.

[33] Surek, G. A. S., Seman, L. O., Stefenon, S. F., Mariani, V. C., and Coelho, L. D. S. 2023. Video-based human activity recognition using deep learning approaches. Sensors 23, 14, 6384.

[34] Turner D., Rodrigues J.M.F., and Rosa M. 2020. Describing People: An Integrated Framework for Human Attributes Classification. In Monteiro J. *et al.* (eds) INCREaSE 2019, 324-336, INCREaSE 2019. pp. 324-336. Springer, Cham. DOI: 10.1007/978-3-030-30938-1_26

[35] Van Haeringen, E. S., Gerritsen, C., and Hindriks, K. V. 2023. Emotion contagion in agent-based simulations of crowds: a systematic review. Autonomous Agents and Multi-Agent Systems 37, 1.

[36] Varghese, E. B., and Thampi, S. M. 2021. Towards the cognitive and psychological perspectives of crowd behaviour: a vision-based analysis. Connection Science 33, 2, 380–405.

[37] Veltmeijer, E. A., Gerritsen, C., and Hindriks, K. V. 2021. Automatic emotion recognition for groups: a review. IEEE Transactions on Affective Computing 14, 1, 89–107.

[38] Wirth, T. D., Dachner, G. C., Rio, K. W., and Warren, W. H. 2023. Is the neighborhood of interaction in human crowds metric, topological, or visual? PNAS nexus 2, 5, pgad118.

[39] Zhang, Y., Qin, L., Ji, R., Zhao, S., Huang, Q., and Luo, J. 2017. Exploring coherent motion patterns via structured trajectory learning for crowd mood modeling. IEEE Transactions on Circuits and Systems for Video Technology 27, 3, 635–648. DOI: 10.1109/TCSVT.2016.2593609.