Unimodal Model for Emotion Detection in Immersive Virtual Learning Environments Using Spatial Analysis of Hands and Head

Jorge E. Velázquez-Cano, Gabriel González-Serna, Leonor Rivera-Rivera, Nimrod González-Franco, Máximo López-Sánchez, and José A. Reyes-Ortiz

***Abstract*—This study introduces a unimodal model for emotion detection in Virtual Reality (VR) environments that depends only on the spatial information from the user's head and hands during interactions within an Immersive Virtual Learning Environment (IVLE). The goal is to eliminate the need for additional sensors, offering an emotional recognition method that can be scaled for multiuser settings. Data on rotation and position from VR devices were collected along with self-reported valence and arousal ratings from 65 participants. Primary and spatial features were extracted, generating mean and median vectors. Random forest regression techniques were then used to predict valence and arousal values in SMOTE augmented data. A paired random pre-augmentation data was used to further test the models in a closer-to-final-implementation scenario. The models achieved accuracies of 70% and 76% for valence prediction using the mean and median vectors, respectively. For arousal, the accuracies were 83% (mean vector) and 87% (median vector). The findings suggest that the median-based approach improves performance, although it involves higher feature dimensionality. This model enables the non-invasive inference of a user's emotional state in VR environments, without cables or extra sensors. This advancement enhances user experience and lowers implementation costs. These results provide a foundation for integrating affective tutors in IVLEs, with potential applications in education and training involving large groups.**

Link to graphical and video abstracts, and to code:
https://latamt.ieeer9.org/index.php/transactions/article/view/10344

***Index Terms*— nonverbal behavior, behavioral measurement, emotion recognition, feature selection, body motion analysis, random forest, Virtual Reality, affecting computing.**

The associate editor coordinating the review of this manuscript and approving it for publication was Ingrid Winkler (*Corresponding author: Jorge Enrique Velázquez Cano*).

This work was supported by the Tecnológico Nacional de México (TecNM), campus Centro Nacional de Investigación y Desarrollo Tecnológico and the Secretaría de Ciencias, Humanidades, Tecnología e Innovación. This research is financed by TecNM.

Jorge Enrique Velázquez Cano, G. González-Serna, N. González-Franco, and M. López-Sánchez are with the Tecnológico Nacional de México, Cuernavaca, Morelos, México (e-mails: d19ce066@cenidet.tecnm.mx, gabriel.gs@cenidet.tecnm.mx, nimrod.gf@cenidet.tecnm.mx, and maximo.ls@cenidet.tecnm.mx).

L. Rivera-Rivera is with the Instituto Nacional de Salud Pública, Cuernavaca, Morelos, México (e-mail: lrivera@insp.mx).

J. A. Reyes-Ortiz is with the Universidad Autónoma Metropolitana, Azcapotzalco, Cdmx, México (e-mail: jaro@azc.uam.mx).

I. Introduction

With the increasing use of artificial intelligence-based tools in tutorial and learning activities, the implementation of affective tutors becomes increasingly relevant. Interaction with these tools is becoming more complete and complex, which makes emotional state an increasingly important factor given its relevance to the learning process [1]. Therefore, emotion detection is a crucial step in implementing virtual affective tutors that enable natural emotional responses between humans and digital avatars. Currently, the main approach to emotion detection involves using multiple devices and complementary systems to process the information. These approaches involve the use of physiological signals [2] [3], such as the electrocardiogram, galvanic skin response, electromyogram, and electroencephalogram [4]; as well as computer vision, which is used for facial recognition [5] or body posture estimation using webcams [6] or Kinect sensors. The use of these additional devices imposes constraints on their widespread implementation (i.e., having as many sensors as users). Also, it interferes with the user's everyday movements and behavior, thereby directly affecting the accuracy of the recordings. The computer vision approach presents challenges related to user privacy, image resolution, intermediate interpretation of joint location, and spatial limitations imposed by the detection devices (e.g., viewing angle, lighting conditions, subject-sensor distance). Equally critical, the constant feeling of being monitored ("sense of being watched") frequently leads users to modify their natural behavior, compromising the ecological validity of the collected emotional data.

Body posture analysis has been addressed in the literature by examining body joints and their relationships [7]. When the target environment requires users to remain seated, body posture is analyzed relative to the upper body, typically using facial analysis. Emotion detection in VR environments is mainly multimodal, using sensors [8], computer vision models to infer expressions from faces occluded by the device [9], or adding electroencephalogram (EEG) sensors to the head-mounted display (HMD) [10] [11]. The multimodal approach is impractical for a massive (i.e., multi-user) implementation because it entails high costs, limited portability, and even usability issues. This paper calls these conditions the "multimodality problem."

Emotion representation is one of the first problems when building a dataset and a ground truth of the emotions sought. Various works use actors and participants who interpret, to the best of their abilities, the stereotypical characteristics of an emotion [12]. However, not everyone expresses their emotions in the same way, and the same subject may do so differently depending on the situation and activity. Furthermore, it is unclear whether an assessment of the actor's emotional state was conducted before the actual portrayal of the intended emotions. The actor's emotional state could influence how they perform other emotions, particularly if they are of opposite valence or arousal.

Emotions are systematic responses of the human body to events and stimuli in any given moment [13], and their categorization has been approached from multiple perspectives [4]. The emotion elicitation or induction process can be passive or active [14] and involves stimuli and situations [15] [16] [17], which can be specific to a group of people or even a single individual. Thus, a particular group of emotions is usually requested, assigning the recorded values to the instruction and not the genuine emotion, or also, the recorded values are labeled with respect to what a group of external observers (experts or not) consider to be the corresponding emotion, but omitting whether the origin of what is represented (i.e. the individual) agrees with that emotion. This paper refers to the two previous conditions as the "A priori labeling problem".

This study aims to address the problems described using a unimodal approach that relies exclusively on data collected from virtual reality (VR) devices during their standard operation. Emotional states are inferred from movement data and associated descriptors (e.g., velocity, acceleration, curvature, symmetry) during interaction with an IVLE. These data are annotated with self-reported valence and arousal values.

A. IVLE ASI

The IVLE "*Pinta tu raya ASI*" (IVLE ASI) was developed in collaboration with the National Institute of Public Health to serve as interactive didactic material for training primary school-aged children in the prevention of child sexual abuse. Developed using Unity Engine and implemented on the Oculus Quest 2 platform, the system has been field-tested with favorable responses. The application consists of eight didactic scenes focused on self-care, featuring interactions that explore the participant's decisions and their consequences. The most critical scene (though neither explicit nor graphically explicit) presents a potential risk situation of child sexual abuse.

Given this condition and those presented in the other scenes, it is essential to monitor the participant's emotional response. If an unfavorable emotional state is induced, timely intervention can return the participant to a neutral or more favorable state, enabling the child to continue the training. In this regard, the work presented here does not address cognitive states related to content retention or comprehension. Given the sensitive nature of the topic, the primary focus is on the emotional reactions elicited by the IVLE ASI material. For this reason, the system aims to extend its capabilities to those of an affective tutor [18], enabling content regulation (interventions for emotional regulation) based on the user's emotional state to promote positive emotions and mitigate negative emotions that may impact the learning process [19]. Tools were integrated to detect and record devices' positions and orientations in real time, along with an emotional self-assessment instrument.

II. Unimodal Detection of Emotional States Using Spatial Information During VR Interactions

A. Participants

A total of 65 volunteers from a primary school (M=36, F=29) participated with prior parental or guardian authorization. Following the interventions and after reviewing the collected data, participants with incomplete or erroneous data or outside the target age group of 8 to 10 years were excluded. The final sample consisted of 58 participants (M=31, F=27). A basic demographic questionnaire was administered to record participants' sex and age. The age distribution was 20%, 80%, and 20% for 8, 9, and 10 years, respectively, as confirmed by a chi-square goodness-of-fit test ($\chi 2 = 1.7471$, $p = 0.4175$). The gender composition was 50% male and 50% female, considered balanced, and was verified by a chi-square goodness-of-fit test ($\chi^2 = 2.2069$, $p = 0.12374$). Participants were not divided into subgroups based on any criteria. No children with special physical or psychological needs were included in the sample. Participants reported minimal or no prior experience with VR devices.

B. Setting

The experiments were conducted with participants using a Meta Quest 2 device, which includes a headset and two controllers. As a standalone device, it requires no additional computing equipment or sensors. Participants were seated in armless chairs to allow unobstructed, unconstrained arm movement, enabling them to adopt a natural, comfortable position during non-interactive scenes. Interactions were presented in two formats: grasping virtual objects positioned directly in front of the user, or pressing three-dimensional buttons located in vertical and horizontal quadrants within the virtual environment, depending on the number of available options, but remaining within arm's reach. Either controller could be used, eliminating the need for specific handedness. These conditions, including the IVLE ASI, define the specific movements required for the interactions as well as the spatial environment in which they take place, together forming the essential context for the expression and detection of emotional states presented in this work. All results presented here were derived directly from experiments conducted under these settings, using the target age group established earlier (Fig. 1).

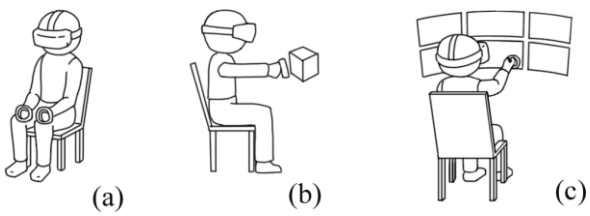

Fig.1. Physical space context, base position (a), and interactions (b and c) with the IVLE ASI.

The experiments followed a straightforward protocol repeated until all participants were covered:

1. A group of six participants was assembled.
2. An introductory and pre-training speech was delivered.
3. VR devices were fitted to each participant, and assistance was provided to initiate interactions.
4. Participants completed the demographic questionnaire in VR.
5. Participants interacted with the IVLE ASI for 15 minutes across eight scenes.
6. After each scene, participants conducted an emotional self-assessment in VR.
7. Upon completing the IVLE ASI, participants completed an exit questionnaire.
8. Voluntary qualitative feedback was requested, and the experiment was concluded.

C. Instruments

A demographic questionnaire was designed and deployed in VR (Fig. 2) to record the participants' sex and age. Each user was assigned a random code to identify specific cases and link VR-based instruments with physical-format instruments (i.e.., paper-and-pencil) while preserving participant anonymity. Emotional self-assessment was conducted using a VR-adapted version of the Self-Assessment Manikin [20], tailored to the target age group and based on studies by [21] and [22], as well as on internal tests with progressive versions. This instrument was named Emoji-SAM VR (Fig. 2).

Fig. 2. Demographic questionnaire interface in VR implemented in the IVLE ASI (Top). Emoji-SAM VR interface implemented in the IVLE ASI (Bottom), displaying the screen corresponding to the valence dimension.

D. Method

The overall structure of the method is presented in Fig. 3. It begins with the collection of location and orientation data at 50 Hz from VR devices used with the IVLE ASI, along with the user's associated self-reported valence and arousal values. These data represent spontaneous natural movements and reactions to what is observed and experienced in the VR environment. Thus, both the information and its label (i.e., expected value) are derived directly from the test subject.

Using location and orientation, primary and spatial features were computed to form the feature set. Orientation angles were recorded in two versions: the raw data provided by the device, ranging from 0° to 359°, and adjusted data derived from a set of rules that allowed negative degrees or values exceeding 359°. This adjustment prevents artifacts in the recording and potential errors in calculating related features when rapid transitions occur between 359° and 0° (Fig. 4).

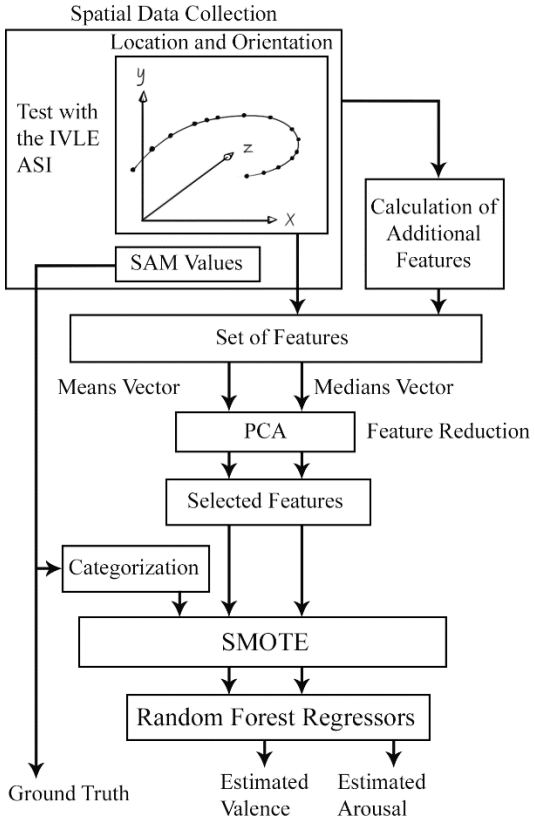

Fig. 3. Diagram of the method structure for detecting emotional states using spatial information.

Primary features included linear and angular velocity and acceleration along each axis. Spatial features included distances between devices (Joint Relative Distances), the triangle area relative to the devices (Joint Relative Triangle Area), the relative angle between devices (Joint Relative Angle), cosine dissimilarity relative to the devices (Joint Relative Cosine Dissimilarity), distal hand symmetry, vertical hand symmetry, delta of distal symmetry and per-axis motion sharpness proxies (one value per axis x, y, z). Related research on posture-based emotion recognition has highlighted the value of geometric and spatiotemporal relationships among body landmarks, primarily joints [23][24][25]. Several of the features explored in this work (Such as Joint Relative Triangle

Area, Joint Relative Angle, Joint Relative Cosine Dissimilarity, and trajectory curvature) draw from this line of research. Notably, curvature was not computed on the overall device trajectory, but separately per-axis curvature proxies were computed to capture abrupt changes in motion along each direction independently:

$$k_i = (v_i * a_i) / v_i^3 \text{ if } v_i \neq 0 \text{ and } a_i \neq 0, \text{ else } k_i = 0 \quad (1)$$

where v_i and a_i are the velocity and acceleration along component i, respectively. When the device is static, no trajectory exists, and curvature is undefined (NaN). To handle edge cases, the rule that $k_i = 0$ whenever v_i or $a_i = 0$, was implemented. This adjustment, may meaningfully influence how curvature contributes to valence and arousal estimation.

By computing curvature proxies separately along each axis, the approach captures directional trajectory sharpness without requiring full 3D vector reconstruction.

Certain features were also included with a relative reference computed using the mean of the recordings from the first second of each scene, identified by the prefix "A" (e.g., AH_ky is the same variable as H_ky but with a different reference frame).

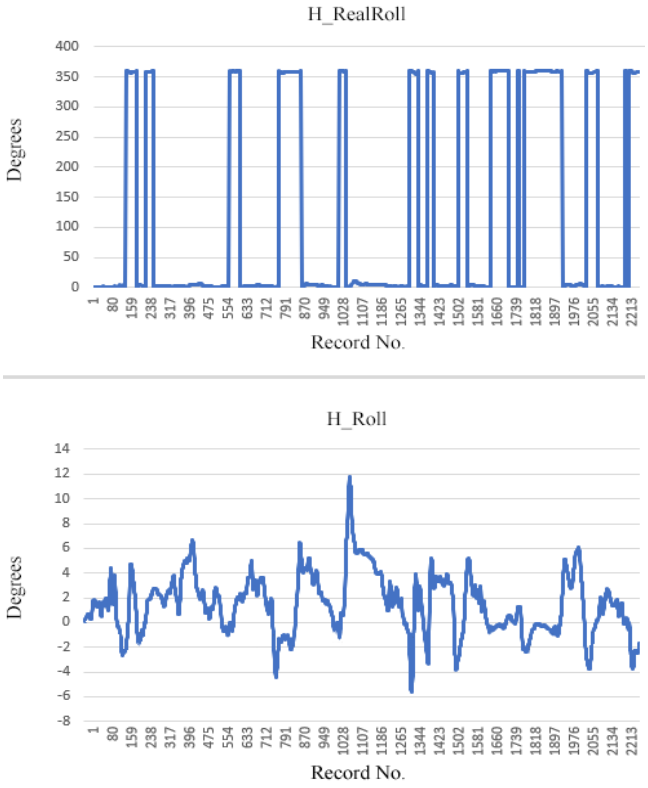

Fig. 4. Raw data recordings from the HMD identified as "Real" (Top), and the same recordings after adjustment to allow degrees above 359 and below 0 (Bottom).

For each user and each scene, feature mean and median vectors were computed, along with the associated valence and arousal values, yielding two data sets. Using principal component analysis, a feature reduction was performed by selecting components that explained 95% of the variance in each data set. Subsequently, the variables constituting these components were identified. The resulting variables form the final vectors (i.e., one for means and one for medians) of selected features. Table I shows the features and indicates the vector to which they belong. The prefixes H, R, and L denote the head, right-hand, and left-hand devices, respectively; the terms Pitch, Roll, and Yaw refer to types of rotation about the respective axes; the letter k indicates trajectory curvature; and the letters x, y, and z specify the axis where applicable.

TABLE I
FEATURES CONTRIBUTING TO THE COMPONENTS THAT EXPLAINED 95% OF THE DATA VARIANCE

Feature	Median	Mean	Feature	Median	Mean
H_RealRoll	*		R_Pitch	*	
H_Roll	*		R_RealYaw	*	
H_RealPitch	*		R_Yaw	*	
H_Pitch	*		H_ky	*	
H_RealYaw	*		AH_ky	*	
H_Yaw	*		H_kz	*	
L_RealYaw	*		L_ky	*	*
L_Yaw	*		AL_ky	*	*
L_RealPitch	*		L_kx	*	
L_Pitch	*		L_kz	*	
R_RealRoll	*		R_kz	*	
R_Roll	*		R_ky	*	
R_RealPitch	*		R_kx		*

To augment and balance the sample count, the dataset was divided into five valence and five arousal categories, with 0.2-interval bins spanning the scale from 0 to 1 (0 = lowest valence/arousal, 1 = highest valence/arousal). Using these categories, SMOTE was applied to each dimension, yielding 1395 observations for valence and 1020 observations for arousal.

E. Valence and Arousal

The Ground Truth and desired output are provided as valence and arousal values. These dimensions are combined into quadrants that distinguish between positive and negative emotional states according to Russell's circumplex model [26].

Valence describes whether an emotion is pleasant or unpleasant, while arousal (also called activation) indicates its energy level. This final statement has proven difficult for the target age group to understand. The SAM refers to the 'size' of the emotion in its classic visual representation (using manikin size to depict arousal), but this can be interpreted as, for example, "slightly happy" or "very happy" (reflecting variations in perceived valence of the same emotion rather than distinct emotions). Additionally, the word "intensity" is often explained using the term "strong" (i.e., how strong is the emotion?) to give the emotion a physical aspect that acts upon the individual. These explanations have little relation to the dimension's objective regarding physiological arousal. Therefore, during the introductory speech, the test subject was asked about their perception of whether an emotion made them feel more energetic (i.e., active) or less energetic (i.e., lethargic).

F. Results of the Regression Techniques

A Random Forest Regressor (RF) was implemented in Python using the Scikit-Learn library. The dataset was split into 80% for training and 20% for testing. GridSearchCV and Optuna were used to optimize hyperparameters. Given the independence of the dimensions, independent RF models were created. Combined with the separate handling of the two central tendency measures, this resulted in two approaches, each with two models. Table II presents the hyperparameters for each model per dimension, and Table III shows the Mean Squared Error (MSE), Mean Absolute Error (MAE), the Root Mean Square Error (RMSE), and the coefficient of determination (R^2) from the training results. The results indicate that using the vector of medians yields better performance, reducing the MSE by more than half in both dimensions.

TABLE II
OPTIMAL HYPERPARAMETERS FOR EACH MODEL, BY APPROACH AND DIMENSION

Approach	Dimension	Estimators	Max depth	Min samples	Min samples leaf	Max features
Mean	Valence	330	37	2	1	2
	Arousal	380	50	2	1	1
Median	Valence	300	17	2	1	4
	Arousal	103	25	2	1	9

TABLE III
EVALUATION OF THE MODELS AFTER TRAINING, BY APPROACH AND DIMENSION

Approach	Dimension	MSE	MAE	RMSE	R2
Mean	Valence	0.039433	0.130334	0.198578	0.589
	Arousal	0.052401	0.165562	0.228914	0.414
Median	Valence	0.015695	0.08216	0.12528	0.833
	Arousal	0.016884	0.089824	0.12993	0.811

An additional test was conducted by taking a random 10% sample of the data from the datasets prior to SMOTE and applying the trained models. This test is significant because it uses pairs of valence and arousal values corresponding to the same trial and user, ensuring a direct correspondence between results, which is the closest approximation to the final implementation. Table IV presents the evaluations of each model, this time showing similar performance across both dimensions.

TABLE IV
EVALUATION OF THE MODELS BY APPROACH AND DIMENSION

Approach	Dimension	MSE	MAE	RMSE	R2
Mean	Valence	0.017241	0.096613	0.131305	0.689
	Arousal	0.018882	0.102374	0.137412	0.586
Median	Valence	0.017107	0.097072	0.130794	0.691
	Arousal	0.012986	0.087536	0.113959	0.715

The same sample and its results were subjected to the previously described categorization to evaluate model performance using confusion matrices and their corresponding metrics, enabling observation of differences between using the mean and median vectors. Fig. 5 compares the confusion matrices for the valence dimension using the median and mean vector models. A low prevalence of classes 1 and 2 is observed in the sample, consistent with findings from the described experiments. Both models perform equally well in category 1, but the mean vector model shows better results in category three and apparently less dispersion in category 5.

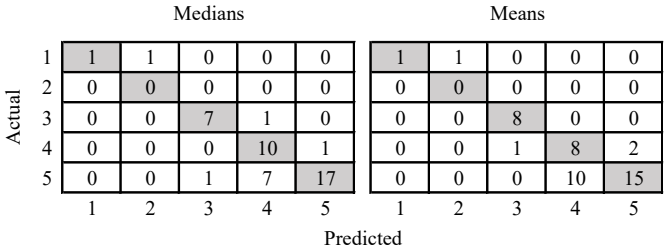

Fig. 5. Confusion matrices of the categorized results for the valence dimension.

The confusion matrix metrics for the valence dimension, along with the accuracy of the respective models, are detailed in Table V. The medians model performs better, particularly for categories 4 and 5.

TABLE V
METRICS OF THE CONFUSION MATRICES FOR THE VALENCE DIMENSION

	Medians				Means			
Classes	Precision	Recall	F1-Score	Support	Precision	Recall	F1-score	Support
1	1	0.5	0.67	2	1	0.5	0.67	2
2	0	0	0	0	0	0	0	0
3	0.88	0.88	0.88	8	0.89	1	0.91	8
4	0.56	0.91	0.69	11	0.44	0.73	0.55	11
5	0.94	0.68	0.79	25	0.88	0.6	0.71	25
Macro avg	0.68	0.59	0.6	46	0.64	0.57	0.57	46
Weighted avg	0.84	0.76	0.78	46	0.78	0.7	0.71	46
Accuracy			0.76	46			0.7	46

The confusion matrices of the results for the arousal dimension are presented in Fig. 6. The superior performance of the median vector model is more evident in these results. It can be observed that classes 2 and 3 have low prevalence, and class 1 is absent in this case.

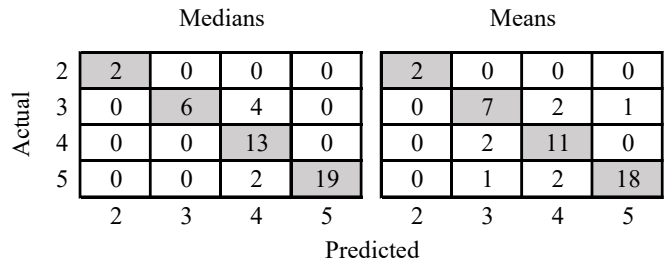

Fig. 6. Confusion matrices of the categorized results for the arousal dimension.

The confusion matrix metrics for the arousal dimension are presented in Table VI. Despite the confusion matrix showing confusion in category 3, the median model achieves a higher F1-score for this category and clearly outperforms the mean vector model on this dimension. Comparing the accuracy of the two vector approaches reveals that both enable better estimation of arousal than of valence.

TABLE VI
METRICS OF THE CONFUSION MATRICES FOR THE AROUSAL DIMENSION

	Medians				Means			
Classes	Precision	Recall	F1-score	Support	Precision	Recall	F1-score	Support
2	1	1	1	2	1	1	1	2
3	1	0.6	0.75	10	0.7	0.7	0.7	10
4	0.68	1	0.81	13	0.73	0.85	0.79	13
5	1	0.9	0.95	21	0.95	0.86	0.9	21
Macro avg	0.92	0.88	0.88	46	0.85	0.85	0.85	46
Weighted avg	0.91	0.87	0.87	46	0.84	0.83	0.83	46
Accuracy			0.87	46			0.83	46

The prevalence of categories 4 and 5 relative to categories 1 and 2 in the sample derives primarily from the IVLE ASI and from participants' perceptions of the VR experience. In addressing the issue of a priori labeling, no attempt was made to induce a specific emotion in any of the scenes to label the collected data as such. The content is presented "as is," and the user's natural reactions and self-assessment shape the dataset. From this, it can be inferred that the interventions induced more positive than negative emotional states. Additionally, while the median vector model clearly performs better, with valence and arousal accuracies 0.06 and 0.04 higher, respectively, the difference is smaller than in the training results. This puts the results into perspective: the mean vector uses only hand-related information, reducing dimensionality and vector size and thereby affecting computational cost.

III. DISCUSSION

This work provides a mechanism for detecting emotional states solely from spatial information collected via VR devices, without additional sensors, constituting a unimodal approach. It also presents a methodology that seeks to avoid the issues identified in this work, namely multimodality and a priori labeling. To the authors' knowledge, an approach with these characteristics, using only head and hand information, has not been previously addressed. In [4], emotion detection through posture is considered a computer vision approach, and in [27]. However, in the later project, affective states are addressed from the perspective of user-reported effort, performance, and perceived frustration, rather than their emotional states.

The advantages of this proposal are: its portability, as it requires no additional devices or equipment beyond the VR devices; non-invasiveness and non-intrusiveness, since the user does not wear any cables from physiological sensors attached to the head or skin, which can lead to greater trust and naturalness by not feeling observed or monitored; comfort, as it leverages the ergonomics designed for the host VR devices; and low computational cost, as, although tests for real-time estimation are still ongoing, it is sufficiently lightweight to execute and provide a response in seconds.

The choice of the feature vector is significant: the mean vector yields notably less favorable initial results compared to the median vector; however, it has a considerably lower dimensionality, using only three features compared to the 25 used by the median vector. It is acknowledged that other features extractable or derivable from the dataset could be explored using frequency analysis techniques, such as the Fast Fourier Transform, or by implementing an alternative regression mechanism, such as a neural network. Preliminary experiments with a neural network model (funnel-like-triple-layer topology) have already been conducted by the authors, yielding promising initial results in the valence dimension (SME=0.008, AME=0.051, R^2=0.91). However, further analysis is required, as current performance may be affected by overfitting and potentially by undiscovered subject-specific characteristics influencing the results.

A. Methodology Considerations

The proposed approach remains non-intrusive within standard VR setups, relying solely on the HMD and controllers as the core platform; consequently, participants reported no nervousness or concerns about additional hardware. The few issues mentioned were instead minor and ergonomics-related (e.g. grip difficulties for a child with smaller hands), without impacting emotional expression or system acceptance. It is worth noting that, while the IVLE ASI inherently supports using only the HMD with natural hand tracking, such an interface can introduce potential noise and imprecision arising from intermediate interpretations of estimated hand positions (e.g., due to occlusion, tracking drift, or variability in gesture detection). For this reason, the study opted to employ physical controllers, prioritizing reliability and precision in capturing movement-based emotional cues within the defined spatial and interaction conditions, at the acceptable cost of minor ergonomic variability across users. This reinforces that the system's non-intrusiveness is conditional upon the immersive VR paradigm itself: the presence of VR devices constitutes the core platform for both interaction and emotion elicitation/detection, rather than an add-on burden. Future refinements in hand-tracking robustness could enable fully controller-free implementations with comparable precision.

Although the methodology is generalizable, the selected variables are suited to the conditions described in the Setting section, as well as potentially to unique conditions of the studied population, which may include socio-cultural aspects not captured by the dataset and beyond the scope of this work. However, these conditions can be adapted to other groups and different VR environments through model retraining using the already identified features, or by re-exploration to determine which features are more relevant in the new setting. This notion aligns with studies such as [27]. Directly reusing information from other studies or experiments to accurately infer a participant's affective state is difficult due to differences in the activities performed, the type of virtual environment, and the target age group.

The handedness of the selected features is currently assumed to be related to the user's dominant hand. To confirm or refute this hypothesis, it is necessary to include a question about handedness in the demographic questionnaire and ensure that left-handed individuals are included.

In this case, the target age group ranges from 8 to 10 years. Studies such as [28] show that different age groups evaluate their emotions differently or identify distinct ranges, either due to a lack of experience or unfamiliarity with the words used to describe them. Additionally, the way emotions are expressed

varies with factors such as gender, age, culture, and context (Picard, 1998, as cited in [29]).

B. Emotional Models and Labeling Choices

There are multiple ways and perspectives for modeling emotions [30]. It is not strange that in computational sciences, there is no unanimously accepted emotional model. However, two broad models have exerted the greatest influence in the field: the discrete or categorical model, primarily represented by Ekman [31] and Plutchik [32], and the continuous or dimensional model, in which Russell [33] proposed his valence-arousal framework.

Ekman's model is grounded in the hypothesis that emotions stem from basic instincts (e.g., fear in response to danger) and are therefore universal across races and cultures. It employs discrete nominal categories (e.g. happy, angry, sad) to represent these emotions. However, emotions rarely manifest in "pure" form; they are typically mixed and complex, and interpretations of basic emotions vary across cultures. This can make it challenging for users to respond when forced to fit their emotion into a predefined set of labels (particularly when the felt emotion is not among the options), either due to limited vocabulary or a developmentally constrained understanding of the emotional spectrum. This situation is part of the broader problem of a priori labeling, in which external agents predetermine which emotions can be recorded or deemed relevant.

Russell's model, by contrast, allows us to bypass the nominal component altogether and provide users with a complete (though not strictly bounded, except perhaps at the extremes) continuous space in which to locate their emotion, regardless of the verbal label they might associate it with. This flexibility is evident in the scatter clouds shown in Fig. 7: clusters of shared values, primarily located in the quadrant of high valence and high arousal, without forcing users into a word they may not know or fully comprehend at that moment. Moreover, given the nature and design of the content (even though no specific emotions were intentionally induced), certain emotions are highly unlikely to occur and therefore do not need to be offered as nominal options a priori.

Although Russell's circumplex model includes a third dimension (Dominance–Submissiveness), the present work deliberately focused solely on the valence and arousal dimensions. This decision was made not only because these two dimensions alone suffice to estimate an emotional state, but also because they significantly simplify the self-reporting process for the target age group (8–10 years). More importantly, preliminary testing during the development of the Emoji-SAM VR instrument revealed that the Dominance dimension represented an overly abstract concept for children in this age range: participants tended to associate the term "dominance" with mastery or understanding of the IVLE ASI didactic content rather than with control or submission over their own emotional state.

During the tests, the nominal labels that Russell associates with regions of the two-dimensional valence-arousal space were also explored. To avoid a priori labeling, an instrument called Nominal Emotional Recording (NER) was designed to identify associations between SAM values and specific self-reported emotions. This instrument was administered in paper format as an exit questionnaire at the end of the IVLE ASI interventions. The scatter plots reveal distinct sectors of valence and arousal where participants appear to locate their emotional state, yet there is a notable lack of agreement in the nominal component, which hinders a direct mapping onto Russell's circumplex model. However, the results from the NER instrument are not yet conclusive and remain under analysis. This pattern may be related to how different age groups assign varying valence values to words with emotional content.

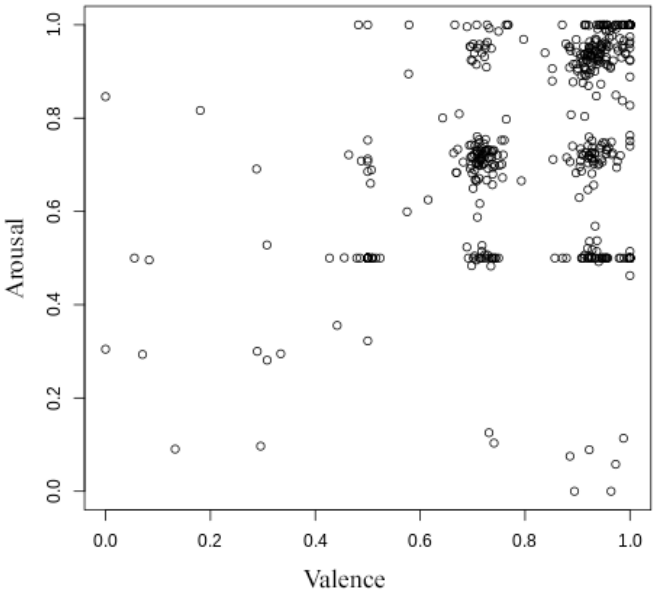

Fig. 7. Distribution of responses from the Emoji SAM VR.

IV. Future Works

[34] Refer to affective features as the geometric and spatial relationships between body points (e.g., hands, head, shoulders) monitored to infer emotional states. They describe postures and movements through the variability of positions over time or within a specific interval. The VR devices collect time-series data, enabling the application of sequential data processing methods to extract temporal features. For this reason, Long Short-Term Memory (LSTM) networks could be considered, as they can store data from specific time steps in a sequence, making them effective at capturing temporal patterns. The implementation of LSTM networks has proven effective in emotion detection [35] [36], as they address the temporal and sequential nature of movements in emotional expression, rather than treating them as fixed moments or single positions. However, emotion detection must also be addressed in the context of its intended use and its role in the process. As of today, implementing techniques such as LSTM networks may not be practical for contexts and applications that require real-time performance. Affective tutoring and content regulation based on emotional state are among such areas. Regarding methodology, other regression mechanisms can be explored, as well as strategies for automatically extracting relevant features, such as autoencoders.

Although there appear to be common elements in how individuals express emotional states, particularly during general actions (e.g., walking), emotions are undoubtedly

individual, and each person expresses them differently through unique gestures and behaviors. Some factors stem from social, familial, and cultural influences. Context is another critical and unavoidable factor, as when performing a task, the tools required to carry it out (if any) or the specific movements needed for its execution form the basis for identifying the peculiarities that distinguish one emotional state from another. A model may perform effectively within the context and activity for which it was trained, and thus its effectiveness is almost certain to decrease if these elements change. Based on this premise, extracting and classifying gestures and microgestures from spatiotemporal information, along with their relationship to the user's emotional state, and training personalized, context-specific models, offers a pathway to more accurate emotion detection.

V. Conclusion

This study demonstrates the potential of a unimodal emotion-detection model that leverages spatial data from head and hand movements within an IVLE, offering a non-invasive, scalable approach for recognizing emotional states in Virtual Reality. By analyzing rotation and location data from VR devices, along with self-assessments and valence and arousal ratings from 8-to-10-year-old participants, we achieved promising accuracies (up to 76% for valence and 87% for arousal) using Random Forest regressors with median-based feature vectors. These results highlight the effectiveness of our approach. However, its performance might be tied to the specific IVLE context, the vector used, and may be influenced by individual factors such as handedness and socio-cultural differences. A broader study group would help to clarify some of these factors. Also, personalized, trained models following the proposed method could be a solution for other IVLEs and contexts. There are still challenges in mapping emotions to established models such as Russell's circumplex, and we are currently developing and testing instruments to address them. Looking ahead, we envision enhancing this model with advanced techniques, such as Long Short-Term Memory networks and autoencoders, to capture temporal patterns and enable real-time emotion detection. By providing an option for non-invasive affective tutoring, our work sets a way for personalized educational experiences, with the potential to transform how emotion detection and regulation are integrated into learning and training environments in VR.

Finally, a key consideration in interpreting these results is the reliance on self-reported ground-truth levels obtained via the SAM scale from children aged 8 to 10 years. Although SAM's pictorial design facilitates intuitive use and has been widely validated in this population, conceptualizing and consistently reporting arousal can pose challenges for some children, potentially introducing variability or noise into the labels. This ground-truth inconsistency may contribute to modest reductions in model reliability (e.g., increased label variability, slightly attenuated performance metrics), particularly on the arousal dimension. These effects remain bounded within the controlled experimental conditions and do not undermine the overall patterns observed. Nonetheless, they highlight the value of complementary validation approaches in future extensions of this work.

References

[1] J. Tan, J. Mao, Y. Jiang, and M. Gao, 'The Influence of Academic Emotions on Learning Effects: A Systematic Review', *International Journal of Environmental Research and Public Health*, vol. 18, no. 18, 2021, doi: 10.3390/ijerph18189678.

[2] F. Alqahtani, S. Katsigiannis, and N. Ramzan, 'Using Wearable Physiological Sensors for Affect-Aware Intelligent Tutoring Systems', *IEEE Sensors Journal*, vol. 21, no. 3, pp. 3366–3378, 2021, doi: 10.1109/JSEN.2020.3023886.

[3] A. F. Bulagang, J. Mountstephens, and J. Teo, 'Multiclass emotion prediction using heart rate and virtual reality stimuli', *Journal of Big Data*, vol. 8, no. 1, p. 12, Jan. 2021, doi: 10.1186/s40537-020-00401-x.

[4] Y. Wang *et al.*, 'A systematic review on affective computing: emotion models, databases, and recent advances', *Information Fusion*, vol. 83–84, pp. 19–52, 2022, doi: 10.1016/j.inffus.2022.03.009.

[5] L. Huang, F. Xie, J. Zhao, S. Shen, W. Guang, and R. Lu, 'Human Emotion Recognition Based on Face and Facial Expression Detection Using Deep Belief Network Under Complicated Backgrounds', *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 34, 05 2020, doi: 10.1142/S0218001420560108.

[6] A. T. S. and R. M. R. Guddeti, 'Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks', *Education and Information Technologies*, vol. 25, no. 2, pp. 1387–1415, Mar. 2020, doi: 10.1007/s10639-019-10004-6.

[7] S. C. Gerdemann, A. Vaish, and R. Hepach, 'Body posture as a measure of emotional valence in young children: a preregistered validation study', *Frontiers in Developmental Psychology*, vol. 3–2025, 2025, doi: 10.3389/fdpys 2025.1536440.

[8] L. B. Hinkle, K. K. Roudposhti, and V. Metsis, 'Physiological Measurement for Emotion Recognition in Virtual Reality', in *2019 2nd International Conference on Data Intelligence and Security (ICDIS)*, 2019, pp. 136–143, doi: 10.1109/ICDIS.2019.00028.

[9] Y. Lin, Y. Lan, and S. Wang, 'A method for evaluating the learning concentration in head-mounted virtual reality interaction', *Virtual Reality*, vol. 27, no. 2, pp. 863–885, June 2023, doi: 10.1007/s10055-022-00689-5.

[10] J. Nam, H. Chung, Y. ah Seong, and H. Lee, 'A New Terrain in HCI: Emotion Recognition Interface using Biometric Data for an Immersive VR Experience', *arXiv [cs.HC]*. 2019, doi: 10.48550/arXiv.1912.01177.

[11] M. Gnacek *et al.*, 'emteqPRO—Fully Integrated Biometric Sensing Array for Non-Invasive Biomedical Research in Virtual Reality', *Frontiers in Virtual Reality*, vol. 3–2022, 2022, doi: 10.3389/frvir.2022.781218.

[12] H. Zhang, P. Yi, R. Liu, and D. Zhou, 'Emotion Recognition from Body Movements with AS-LSTM', in *2021 IEEE 7th International Conference on Virtual Reality (ICVR)*, 2021, pp. 26–32, doi: 10.1109/ICVR51878.2021.9483833.

[13] K. R. Scherer, 'What are emotions? And how can they be measured?', *Social Science Information*, vol. 44, no. 4, pp. 695–729, 2005, doi: 10.1177/0539018405058216.

[14] R. Somarathna, T. Bednarz, and G. Mohammadi, 'Virtual Reality for Emotion Elicitation – A Review', *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 2626–2645, 2023, doi: 10.1109/TAFFC.2022.3181053.

[15] J. Reichenberger, M. Pfaller, and A. Mühlberger, 'Gaze Behavior in Social Fear Conditioning: An Eye-Tracking Study in Virtual

Reality', *Frontiers in Psychology*, vol. 11–2020, 2020, doi: 10.3389/fpsyg.2020.00035.
[16] I. Kritikos, G. Tzannetos, C. Zoitaki, S. Poulopoulou, and D. Koutsouris, 'Anxiety detection from Electrodermal Activity Sensor with movement & interaction during Virtual Reality Simulation', in *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*, 2019, pp. 571–576, doi: 10.1109/NER.2019.8717170.
[17] L. Petrescu *et al.*, 'Integrating Biosignals Measurement in Virtual Reality Environments for Anxiety Detection', *Sensors*, vol. 20, no. 24, 2020, doi: 10.3390/s20247088.
[18] S. Petrovica, A. Anohina-Naumeca, and H. K. Ekenel, 'Emotion Recognition in Affective Tutoring Systems: Collection of Ground-truth Data', *Procedia Computer Science*, vol. 104, pp. 437–444, 2017, doi: 10.1016/j.procs.2017.01.157.
[19] M. A. Hasan, N. F. M. Noor, S. S. B. A. Rahman, and M. M. Rahman, 'The Transition From Intelligent to Affective Tutoring System: A Review and Open Issues', *IEEE Access*, vol. 8, pp. 204612–204638, 2020, doi: 10.1109/ACCESS.2020.3036990.
[20] M. M. Bradley and P. J. Lang, 'Measuring emotion: The self-assessment manikin and the semantic differential', *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994, doi: 10.1016/0005-7916(94)90063-9.
[21] A. Betella and P. F. M. J. Verschure, 'The Affective Slider: A Digital Self-Assessment Scale for the Measurement of Human Emotions', *PLOS ONE*, vol. 11, no. 2, pp. 1–11, 02 2016, doi: 10.1371/journal.pone.0148037.
[22] E. C. S. Hayashi, J. E. G. Posada, V. R. M. L. Maike, and M. C. C. Baranauskas, 'Exploring new formats of the Self-Assessment Manikin in the design with children', in *Proceedings of the 15th Brazilian Symposium on Human Factors in Computing Systems*, São Paulo, Brazil, 2016, doi: 10.1145/3033701.3033728.
[23] Y. Bhatia, A. H. Bari, G.-S. J. Hsu, and M. Gavrilova, 'Motion Capture Sensor-Based Emotion Recognition Using a Bi-Modular Sequential Neural Network', *Sensors,* vol. 22, no. 1, 2022. doi:10.3390/s22010403.
[24] A. S. M. H. Bari and M. L. Gavrilova, 'Artificial Neural Network Based Gait Recognition Using Kinect Sensor', *IEEE Access*, vol. 7, pp. 162708–162722, 2019, doi: 10.1109/ACCESS.2019.2952065.
[25] S. Piana, A. Staglianò, F. Odone, and A. Camurri, 'Adaptive Body Gesture Representation for Automatic Emotion Recognition', *ACM Trans. Interact. Intell*. Syst., vol. 6, no. 1, Mar. 2016. doi:10.1145/2818740.
[26] J. A. Russell, 'Core affect and the psychological construction of emotion', *Psychological Review,* vol. 110, no. 1, pp. 145–172, 2003, doi: 10.1037/0033-295X.110.1.145.
[27] V. Holzwarth *et al.*, 'Towards estimating affective states in Virtual Reality based on behavioral data', *Virtual Reality*, vol. 25, no. 4, pp. 1139–1152, Dec. 2021, doi: 10.1007/s10055-021-00518-1.
[28] L. Sabater, M. Guasch, P. Ferré, I. Fraga, and J. A. Hinojosa, 'Spanish affective normative data for 1,406 words rated by children and adolescents (SANDchild)', *Behavior Research Methods*, vol. 52, no. 5, pp. 1939–1950, Oct. 2020, doi: 10.3758/s13428-020-01377-5.
[29] A. Kleinsmith and N. Bianchi-Berthouze, 'Affective Body Expression Perception and Recognition: A Survey', *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 15–33, 2013, doi: 10.1109/T-AFFC.2012.16.
[30] J. Tracy and D. Randles, 'Four Models of Basic Emotions: A Review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt', *Emotion Review*, vol. 3, pp. 397–405, 09 2011. doi: 10.1177/1754073911410747.
[31] P. Ekman, 'An argument for basic emotions', *Cognition and Emotion*, vol. 6, no. 3–4, pp. 169–200, 1992. doi:10.1080/02699939208411068.
[32] R. Plutchik, 'A psychoevolutionary theory of emotions', *Social Science Information*, vol. 21, no. 4–5, pp. 529–553, 1982. doi:10.1177/053901882021004003.
[33] J. Russell, A Circumplex Model of Affect, *J. Pers. Soc. Psychol*. 39 1980 1161–1178. doi:10.1037/h0077714.
[34] T. Randhavane, U. Bhattacharya, K. Kapsaskis, K. Gray, A. Bera, and D. Manocha, 'Identifying Emotions from Walking using Affective and Deep Features', *arXiv [cs.CV]*. 2020, doi: 10.48550/arXiv.1906.11884.
[35] Z. Zhang, J. M. Fort, and L. Giménez Mateu, 'Facial expression recognition in virtual reality environments: challenges and opportunities', *Frontiers in Psychology*, vol. 14–2023, 2023, doi: 10.3389/fpsyg.. 2023.1280136.
[36] C. Wang, T. S. Kumar, W. De Raedt, G. Camps, H. Hallez, and B. Vanrumste, 'Drinking Gesture Detection Using Wrist-Worn IMU Sensors with Multi-Stage Temporal Convolutional Network in Free-Living Environments', in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2022, pp. 1778–1782, doi: 10.1109/EMBC48229.2022.9871817.

Jorge Velázquez-Cano received his Bachelor's Degree in Cybernetic Engineering and Computational Systems from La Salle Cuernavaca University, Cuernavaca, Morelos, Mexico, in 2007. The Master's degree in Computer Sciences, in 2021, from the National Center for Technological Research and Development (TecNM/CENIDET), Cuernavaca, Morelos, Mexico. His studies are funded by the Secretariat of Science, Humanities, Technology and Innovation. (SECIHTI). He is currently a Ph.D. student at TecNM/CENIDET. His research fields include Affective Computing, Human-Computer Interaction in VR, and Emotion Recognition.

Gabriel González-Serna is a Computer Systems Engineer who graduated from the Technological Institute of Acapulco (TecNM/ITA) in Acapulco, Guerrero, Mexico, in 1992. He earned his Master's Degree in Computer Science from the National Center for Technological Research and Development (TecNM/CENIDET) in Cuernavaca, Morelos, Mexico, in 1995. In 2006, he obtained his Ph.D. in Computer Science from the Computer Research Center of the National Polytechnic Institute (CIC-IPN) in Mexico City (CDMX). González-Serna joined TecNM/CENIDET as a researcher in Intelligent Hybrid Systems in 1992. Since 1995, he has been a Professor-Researcher in the Department of Computational Sciences at TecNM/CENIDET. His research interests include human-computer interaction, affective computing, and user experience (UX).

Leonor Rivera-Rivera is a Medical Surgeon who graduated from the Faculty of Medicine of the Autonomous University of Nayarit. She received her Master's Degree in Health Sciences in Reproductive Health at the National Institute of Public Health, Cuernavaca, Morelos, Mexico, and her Ph.D. in Psychology and Health from the Faculty of Psychology of the National Autonomous University of Mexico, Coyoacán, Ciudad de México, Mexico. She is a researcher in Medical Sciences at the National Institute of Public Health of Mexico. She has led several research projects, and her scientific publications cover topics such as violence against women, dating violence, depressive symptomatology, suicidal behavior, sexual abuse, reproductive health, mental health, and addiction issues. Dr. Rivera is a member of the National System of Researchers (SNI II) of the Secretariat of Science, Humanities, Technology and Innovation. (SECIHTI).

Nimrod González-Franco, received a Ph.D. degree in computer science in 2017 from the National Center for Technological Research and Development (TecNM/CENIDET), Cuernavaca, Morelos, Mexico. He joined TecNM/CENIDET in Cuernavaca, Mexico, as a research professor in the Intelligent Hybrid Systems area in 2019. His research areas include brain-computer interface systems and machine learning.

José Alejandro Reyes-Ortiz received his Master's degree in Computer Science from the National Center for Research and Technological Development in Cuernavaca, Morelos, Mexico (2008), and his Ph.D. in Computer Science from the same center (2013). He is a full-time professor in the Systems Department at the Metropolitan Autonomous University, Azcapotzalco. He is currently the Head of the Systems Department at the Metropolitan Autonomous University, Azcapotzalco. His research interests include knowledge representation, natural language processing, machine learning, and deep learning.

Máximo López-Sánchez is an Industrial Engineer who graduated from the Technological Institute of Zacatepec (TecNM/ITZ) in Zacatepec, Morelos, Mexico, in 1975. He earned his Master's Degree in Computer Sciences from the National Center for Technological Research and Development (TecNM/CENIDET) in Cuernavaca, Morelos, Mexico, in 1994. He obtained his Ph.D. in Computer Research Center from the National Polytechnic Institute (CIC-IPN) in CDMX, Mexico, in 2004. He is a Professor-Researcher in the Department of Computational Sciences at TecNM/CENIDET. His research interests are software system modeling and real-time systems.