Exploring Spontaneous Emotion Recognition from Face Tracking in Head-Mounted Displays

Julia Domingo-Ajenjo[1,*], Almudena Palacios-Ibañez[2], Alberto Altozano[1], Maria Eleonora Minissi[1], Manuel Contero[1], Mariano Alcañiz[1] and Javier Marín-Morales[1]

[1]University Research Institute of Human-Centered Technologies, Universitat Politècnica de València, Valencia, 46021, Spain
[2]Universitat Jaume I, Castellón de la Plana, Comunitat Valenciana, Spain

Abstract

Head-mounted displays (HMDs) are becoming valuable in fields such as education, gaming, and digital health, where understanding user emotion is increasingly important. However, HMD-based emotion recognition remains challenging. First, headset position constrains the mobility of key facial regions. Furthermore, generating reliable ground truth labels is challenging since external annotators cannot see the user's actual face for accurate emotion assessment. We study facial emotion recognition from on-board face-tracking signals recorded with a Meta Quest Pro while 49 participants viewed affective images and videos in a non-acted protocol. Additionally, an external labeling procedure was conducted using synthetically recreated participants' faces. Using nested cross-validation and One-Versus-All XGBoost classifiers, we obtain strong classification performance. Happiness achieved the highest scores (balanced accuracy = 0.94), while surprise, sadness, disgust, and neutral reached balanced accuracies of 0.77–0.80. These findings demonstrate that annotations derived from synthetic facial reconstructions can serve as a viable ground truth. While certain limitations remain, the results demonstrate that the models are capable of effectively recognizing emotions.

Keywords

Head-Mounted Displays, Face Tracking, Emotion Recognition, Affective Computing, Explainable AI

1. Introduction

Head-Mounted Displays (HMDs) enhance users' sense of presence and immersion in virtual environments, which has led to their growing adoption across multiple domains. In education, they have been shown to boost student motivation, support collaborative learning, and foster autonomous strategies and self-efficacy [1, 2]. In gaming, HMDs enable more natural and engaging experiences, while serious games that employ them show strong potential for assessing and rehabilitating motor dysfunctions [3, 4]. In digital health and psychology, HMD-based VR allows controlled anxiety induction for exposure therapies and enables real-time monitoring of mental workload during cognitive tasks [5, 6].

As VR adoption grows, emotion recognition has become essential for personalizing experiences, boosting engagement, and supporting affective computing [7]. It has been shown to enhance immersion in gaming [8] and enable emotional assessment in psychological contexts [9]. Emotional responses can be captured via a wide range of signals, including physiological measures such as EEG or ECG [7, 10] which offer valuable insights into the body's reactions to emotional states. However, these methods are often costly and have low scalability.

Conversely, behavioral sensors provide a more ecological and scalable approach. Speech- or movement-based methods, which integrate more easily and collect data with minimal user effort, face strong limitations when participants are silent or motionless. On the other hand, facial expression analysis, rooted in psychology and informed by the Facial Action Coding System (FACS) [11, 12], which describes facial muscle movements as Action Units (AUs), provides non-invasive real-time detection of emotional cues through micro-expressions. Previous work has shown that face tracking can be

SAXR Workshop at CHI 2026: ACM Conference on Human Factors in Computing Systems, Barcelona, Spain, April 13–17, 2026
*Corresponding author.
jdomaje@htech.upv.es (J. Domingo-Ajenjo); almudena.palacios@uji.es (A. Palacios-Ibañez); aaltfer@htech.upv.es (A. Altozano); meminiss@htech.upv.es (M. E. Minissi); mcontero@upv.es (M. Contero); malcaniz@htech.upv.es (M. Alcañiz); jamarmo@htech.upv.es (J. Marín-Morales)

effectively applied in emotion recognition [13, 14]. However, a current challenge is the lack of studies developing or proposing methodologies to model facial expressions specifically adapted to HMDs, underscoring the need to carefully consider how emotional expressions are elicited in experimental research.

There are two main approaches to eliciting emotional facial expressions in research: spontaneous and posed. Spontaneous methods evoke genuine responses through stimuli such as evocative images or memory recall [15], whereas posed methods require participants to deliberately display expressions without necessarily experiencing the emotion [16]. Both approaches have limitations related to human bias, ecological validity, and variability. Posed expressions can be exaggerated or fail to reflect genuine affect, while spontaneous elicitation may be ineffective if the target emotion is not fully experienced. Given these challenges in reliably eliciting genuine affect, an external ground truth becomes indispensable, as shown in prior studies [17, 18].

However, in HMD-based emotion elicitation, this requirement becomes inherently difficult to satisfy due to two fundamental challenges. First, the headset physically occludes large portions of the face, particularly the eyes and eyebrows, which convey critical affective cues [19]. This occlusion restricts the information available to computational models, directly impairing automated emotion recognition. Second, the HMD's facial coverage hinders external annotation, as evaluators cannot observe the upper facial region, making reliable ground truth labeling difficult. This limitation constitutes a significant methodological challenge, particularly given that classical computer vision models have traditionally been developed and trained on fully visible facial data [20].

To address these challenges we introduce a methodology that integrates spontaneous audiovisual emotion elicitation with external annotations derived from synthetic facial reconstructions. Using facial-gesture signals captured by the Meta Quest Pro HMD, we generate avatar-based facial renderings that restore the full facial configuration and allow independent evaluators to observe and label expressions without the visual constraints imposed by the device. By removing these occlusion-related limitations, the proposed reconstruction enables third-party assessment that would otherwise be unachievable, effectively operationalizing an external ground truth while preserving ecological capture conditions. This approach corroborates spontaneous emotional responses and increases the reliability and consistency of the resulting annotations. Furthermore, relying exclusively on data obtained from the Meta Quest Pro ensures ecological validity and scalability, avoiding the need for complex multimodal pipelines while supporting robust One-Versus-All (OvA) emotion classification adaptable to real-world individual and social contexts.

2. Materials and Methods

2.1. Participants

For this exploratory study, 49 native Spanish-speaking participants were recruited (21 female, 28 male; age 20–55, M = 30.7, SD = 8.4). To minimize biases in emotion elicitation, participants were screened for depressive symptoms and state anxiety using the PHQ-9 [21] and STAI-Y1 [22]. Only participants scoring below 9 on the PHQ-9 and below 38 on the STAI-Y1, which correspond to moderate symptom and level thresholds, were included in the study. The study received ethics approval from the Universitat Politècnica de València, and all participants provided informed consent.

2.2. Experimentation and data collection

To elicit emotional responses, participants were presented with either affective images or short video clips, corresponding to two widely employed emotion-elicitation modalities. All stimuli were delivered within an immersive VR environment. Accordingly, we utilized the IAPS database [23] to provide validated emotional images and developed a custom set of emotional video clips. The inclusion of both stimulus types enabled broader coverage of expressive and affective variability, as different media formats are known to evoke distinct emotional responses.

The experimental procedure comprised two elicitation-based methods:

2.2.1. Experiment 1. Image-Based Emotion Elicitation

This experiment focused on eliciting, through static images, six primary emotions: happiness, sadness, disgust, anger, surprise, and fear, along with a neutral condition. For each target emotion, twelve stimuli were selected from the IAPS database. Twenty-six participants (15 male, 11 female; M = 28.3, SD = 8.3) completed consent before viewing blocks of six emotional images and one neutral image. Each stimulus was displayed for 5 seconds while the VR headset recorded their facial expressions continuously.

2.2.2. Experiment 2: Video-Based Emotion Elicitation

This experiment used dynamic audiovisual stimuli to induce the same six emotions, excluding neutral. Twelve internet-sourced video clips were initially selected as candidate stimuli for the target emotions. They were validated by twelve independent raters, and for each emotion, the clip with the highest agreement was retained as the final stimulus. The selected videos ranged from 40 seconds to 2 minutes, for a total experimental duration of approximately 8 minutes and 50 seconds. Twenty-three participants (13 male, 10 female; M = 32.4, SD = 7.0) completed consent before watching the six validated videos in randomized order. Each clip was followed by a 10-second rest and self-report ratings, while the VR headset continuously tracked facial movements.

2.3. Spontaneous Emotional Face Tracking Dataset

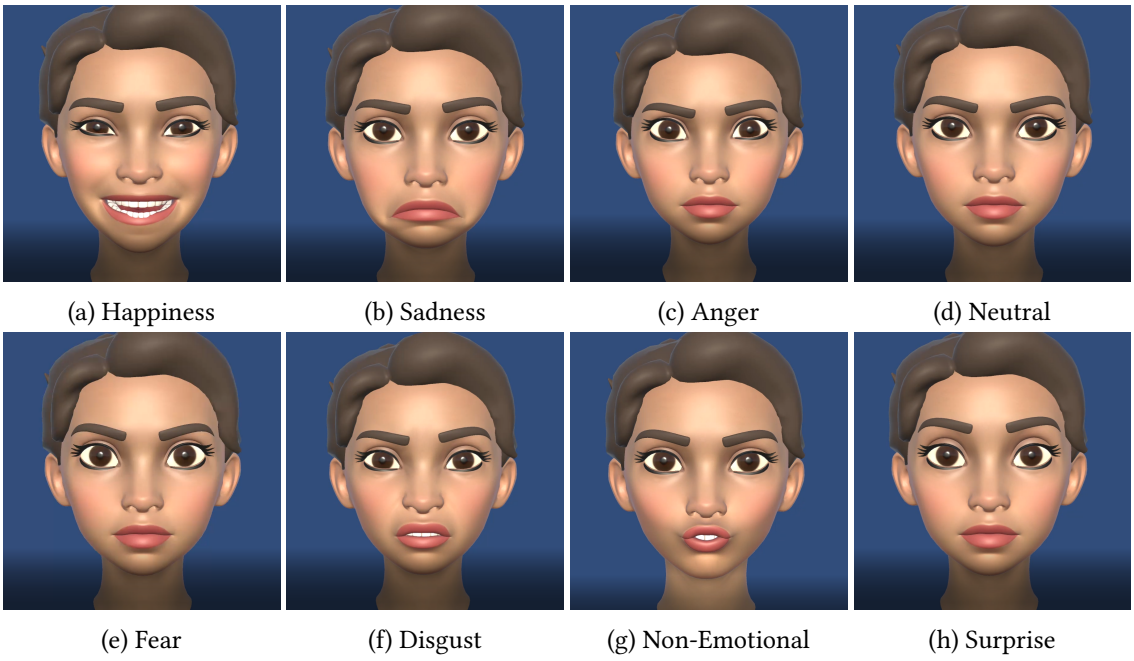
The Meta Quest Pro headset captures facial movements through internal sensors mapped to 63 FACS-based blendshapes, sampled at 50 Hz with values ranging from 0 to 1. Blink-related blendshapes were removed during preprocessing.

Although each stimulus includes theoretical emotion labels, the corresponding facial expressions are not always consistently present, as emotional responses may only emerge at specific moments. To address this limitation and obtain more accurate annotations of facial expressions, we performed an external labeling.

The first step was to distinguish moments when participants exhibited emotional responses from those with neutral expressions. To this end, the Mahalanobis distance was computed for all samples, and only the top 15% most distinctive samples were retained as emotionally expressive. The remaining 85% were excluded from manual labeling, as they primarily reflected low-intensity or neutral expressions. The selected samples were grouped into continuous segments, with segments shorter than 1 second discarded. This procedure resulted in a total of 6,279 one-second expressive facial segments, of which 2,046 corresponded to image-based elicitation and 4,233 to video-based elicitation.

The 6,279 segments were manually labeled by an expert evaluator into the categories “happy”, “sad”, “neutral”, “surprise”, “disgust”, “anger”, and “fear”. While this process captures the most distinctive facial expressions, some segments could include movements not directly related to emotions, such as sneezes, yawns, or headset adjustments. To account for these, an additional “non-emotional” category was introduced. To facilitate this, GIF visualizations of each segment were generated in Unity by animating a head model with the recorded FACS values. The avatar used for this purpose was developed by Meta and is specifically optimized for high-fidelity blendshape and eye-tracking transmission from the Meta Quest Pro headset, ensuring accurate reproduction of the participants’ facial movement (Fig. 1). A custom labeling application was developed to streamline annotation, providing buttons for the six emotions, neutral, and a non-emotional category for those expressions corresponding to noise or emotional states not included in the target emotion set (Fig. 2).

The final dataset included the curated face-tracking segments, participant IDs, and manually assigned emotion labels. The distribution of labeled samples was as follows: 2,356 neutral, 1,207 non-emotional, 620 surprise, 536 sadness, 492 happiness, 446 disgust, 362 anger and 260 fear. To address class imbalance, random selection reduced the predominance of neutral and non-emotional samples. Additionally, a subset of randomized data segments previously excluded using the Mahalanobis distance threshold was



(a) Happiness (b) Sadness (c) Anger (d) Neutral

(e) Fear (f) Disgust (g) Non-Emotional (h) Surprise

Figure 1: Specific frames for emotions obtained from head model implemented in Unity.

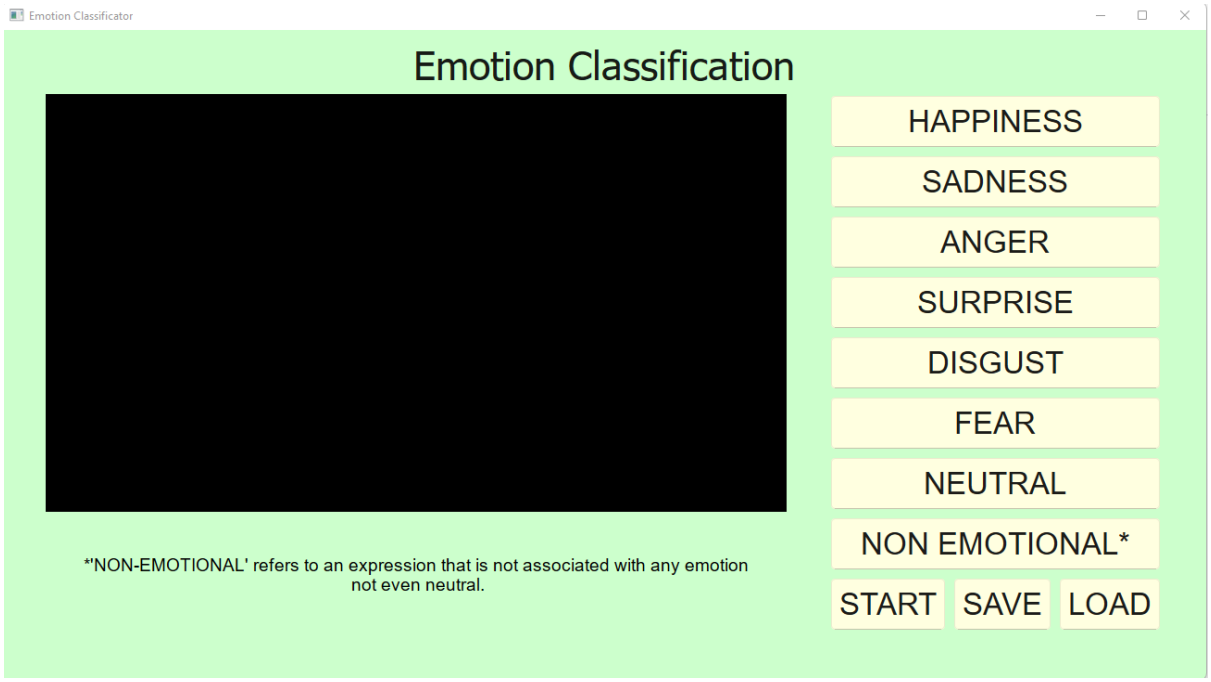


Figure 2: Application developed for emotion classification.

resampled into continuous 1-second windows and incorporated under the neutral label, as their low expressiveness aligned with the operational definition of neutrality. The resulting dataset contained 400 samples each for neutral and non-emotional classes, including 150 neutral segments originally discarded by the Mahalanobis criterion.

2.4. Data modeling and Validation

First, to prepare the data for modeling, each face-tracking segment, lasting 1 second (50 frames, with each frame representing 0.02 seconds), was reduced to a feature vector by computing the maximum

activation value of each blendshape, representing the most expressive point. Additionally, to account for individual baseline differences, we applied subject-wise normalization by subtracting from each feature vector each participant's average blendshapes during a blank-image viewing period.

As for our modeling strategy, we implemented a nested cross-validation scheme using OvA XGBoost classifiers for each emotion, including the neutral and non-emotional classes. In the outer loop, participants were split into training and test sets on a subject-wise basis to ensure that no individual appeared in both. For classification, each OvA model generated a probability score for its target emotion, with the class with the maximum probability assigned as the predicted label. The inner loop performed hyperparameter tuning and class-weight balancing. The hyperparameter tuning was carried out using the Hyperopt library. The values of n_estimators and learning_rate were fixed at 300 and 0.01, respectively. The optimization focused on the following parameters: max_depth ranging from 3 to 10 with a step size of 2; min_child_weight ranging from 1 to 6 with a step size of 2; gamma ranging from 0.1 to 2; reg_alpha ranging from 1 to 5 with a step size of 1; and colsample_bytree ranging from 0.5 to 0.9.

After model training, performance was evaluated using balanced accuracy, F1-scores, True Positive Rate (TPR), True Negative Rate (TNR) and ROC/AUC score for each classifier, under a binary framework where each model distinguished its target emotion from all others. Evaluation metrics were computed by fold aggregation.

2.5. SHAP Explanation Values

To obtain interpretable insights into the model's decision-making process, SHAP values were computed for each emotion class to identify the blendshape coefficients that contributed most strongly to the predicted outcomes. This method allows us to examine whether the learned decision patterns are consistent with FACS-based facial movements and to confirm that the classifiers rely on meaningful expressive cues rather than artefacts or noise. In this way, SHAP contributes to model transparency and supports the detection of potential sources of bias.

Because SHAP explanations depend on the specific model configuration from which they are derived, it was essential to compute them on models operating under their best-performing settings. For this reason, all classifiers were re-trained on the full dataset using 5-fold cross-validation for hyperparameter optimization. The hyperparameters yielding the highest accuracy were selected so that the subsequent SHAP analyses would reflect the models' optimal behavior.

After the optimal classifiers were obtained, each OvA estimator was stored independently, allowing SHAP values to be extracted for each emotion class separately. Explanations were computed using all available training samples to maximize coverage of the data distribution and to better characterize how each blendshape contributed across the full range of expressive variability.

Finally, SHAP summary plots were generated to visualize both feature importance and feature influence. These plots display the distribution of SHAP values using color gradients (red for higher feature activation, blue for lower), enabling identification of the five blendshapes with the strongest impact on each model's predictions and illustrating how their activation levels shaped the final classification decisions.

3. Results

In terms of model performance, the results obtained are shown in Table 1. Overall, the results indicate that happiness was the most accurately classified emotion, achieving a balanced accuracy of 0.94 and an F1-score of 0.85, as reported in Table 1. Surprise, sadness, disgust, and neutral also demonstrated relatively high classification performance, with balanced accuracies of 0.79, 0.80, 0.79, and 0.77; and F1-scores of 0.66, 0.65, 0.64, and 0.53, respectively. In contrast, fear and the non-emotional categories, despite attaining balanced accuracies of 0.72 and 0.68, exhibited lower F1-scores of 0.48 and 0.45, respectively. Anger was the most poorly classified emotion, with a balanced accuracy of 0.55 and an F1-score of 0.19. Overall, the least represented emotions, anger and fear, showed the lowest classification performance,

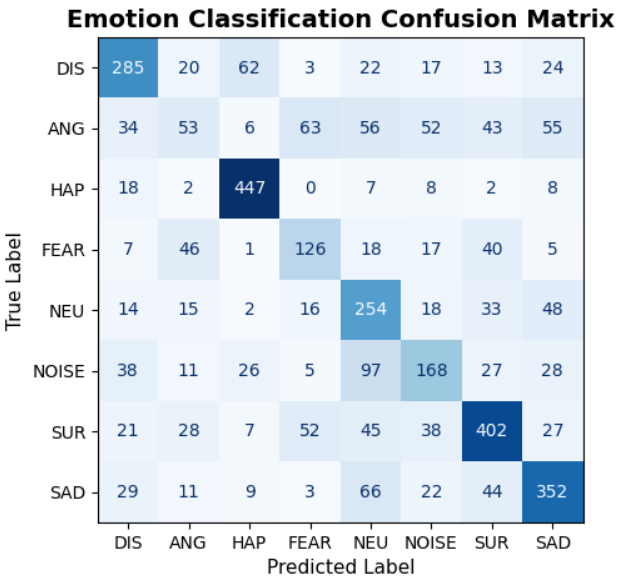


Figure 3: Confusion Matrix implemented for emotion classification task. Disgust (DIS), Anger (ANG), Happiness (HAP), Neutral (NEU), Non-Emotional (NOISE), Surprise (SUR), Sadness (SAD).

with the notable exception of the non-emotional class, which also displayed low performance despite its higher representation in the dataset.

Also, a confusion matrix was constructed to evaluate how individual OvA models classified emotions. The results of all folds were aggregated into a single confusion matrix (Fig. 3). It indicates that happiness achieved the highest prediction performance among all classes, correctly identifying 447 of 492 test samples. The model also demonstrated relatively strong performance for surprise, sadness, disgust, and neutral, correctly classifying 402/620, 352/536, 285/446, and 254/400 samples, respectively. These results suggest that the model can reliably recognize these emotions and the neutral class, although some misclassifications occurred: surprise was primarily misclassified as fear, disgust as happiness, sadness as neutral, and neutral as sadness. In contrast, the classification performance for fear and non-emotional samples was considerably lower, with only 126/260 and 168/400 samples correctly predicted, respectively. Misclassifications contributed to this reduced performance, with fear often labeled as anger or surprise, and non-emotional samples predominantly misclassified as neutral. The anger class exhibited the weakest overall performance, with only 53 of 362 samples correctly classified. Anger instances were misclassified across nearly all other categories except happiness, most frequently as fear, neutral, or sadness.

Table 1
Balanced accuracies (Bal. Acc.), F1-scores, True Positive Rates (TPR), True Negative Rates (TNR) and ROC-AUC score obtained for each OvA model.

Emotion	Bal. Acc.	F-1 scores	TPR	TNR	ROC-AUC
Happiness	0.94	0.85	0.91	0.96	0.94
Surprise	0.79	0.66	0.64	0.95	0.79
Sadness	0.80	0.65	0.66	0.93	0.79
Disgust	0.79	0.64	0.64	0.95	0.79
Neutral	0.77	0.53	0.65	0.93	0.77
Fear	0.72	0.48	0.48	0.96	0.72
Non-Emotional	0.68	0.45	0.42	0.94	0.68
Anger	0.55	0.19	0.15	0.96	0.55

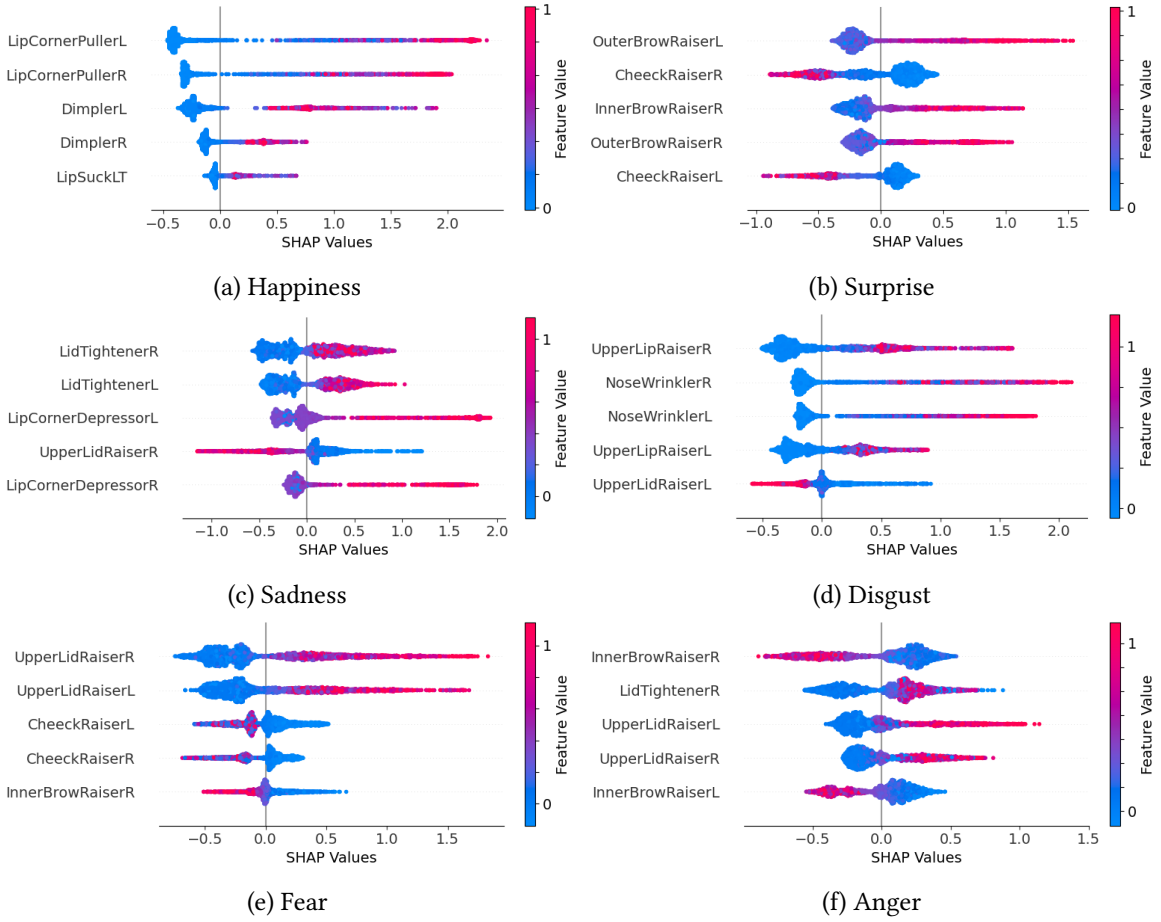


Figure 4: SHAP summary plots for each OvA classifier.

SHAP Explanation Values for each model related to emotions are represented in Summary Plots (Fig. 4). The values computed for each OvA estimator align with FACS theory [11]. For happiness, the most influential features correspond to smile-related blendshapes, and their positive SHAP values indicate that higher activation of these blendshapes increases the likelihood of the happiness class. For surprise, blendshapes associated with widened eyes and raised eyebrows show the greatest positive contribution to the model's decision. In the case of sadness, the most relevant features include blendshapes related to lip depressing and partial eye closure, which are commonly associated with crying or sad expressions; higher values of these blendshapes positively influence sadness classification. For disgust, blendshapes linked to facial wrinkling exhibit the strongest impact. Regarding fear, blendshapes involving upper eyelid raising contribute significantly and are consistent with FACS expectations; however, other fear-related blendshapes, like InnerBrowRaiser, show SHAP contributions opposite to those theoretically expected. Finally, for anger, the most relevant blendshapes correspond to eyebrow lowering and tightened eyelids, which aligns with FACS. This is also consistent with patterns observed in misclassifications, as anger is most frequently confused with fear and sadness, which share the same AUs.

4. Discussion

In this study, we propose a novel methodology for facial emotion recognition in HMD experimental settings. The approach combines spontaneous emotion elicitation with externally annotated labels derived from synthetically reconstructed facial representations, ensuring high-quality training data that closely reflects naturalistic emotional behavior. Furthermore, the methodology is validated through explainable machine learning techniques, providing interpretable insights into the model's decision-

making process.

To implement this approach, we employed emotion elicitation using validated images and ad-hoc selected videos. Among these, videos proved to be the most informative, generating the largest amount of facial expression data (4,233 video-based GIFs vs 2,046 image-based). To maximize the captured information, both types of stimuli were combined into a single dataset, ensuring broad coverage of facial expressions. Nevertheless, incorporating additional elicitation methods, such as interactive virtual agents or serious games, could further enhance the diversity and quantity of captured expressions. Importantly, the primary value of this dataset lies in the spontaneity of the elicited emotions, which minimizes posed biases and preserves the natural dynamics of facial expressions.

In traditional emotion-recognition studies, models rely on full-face data [8, 20], which is not available in HMD settings due to occlusion. Our use of Meta Avatar reconstruction mitigates this limitation by recovering the obscured facial regions with high anatomical fidelity, enabling more reliable external annotation of spontaneous expressions. This approach reduces dependence on self-reports and allows for higher-quality ground truth labels, ultimately improving the robustness of emotion analysis in immersive VR environments.

Despite being trained exclusively on spontaneous expressions, the resulting pipeline demonstrated strong predictive performance. Happiness achieved the highest classification accuracy, with only a few instances misclassified, indicating that it forms a well-defined and robust category. Similarly, surprise, sadness, and disgust exhibited relatively strong performance. In contrast, lower classification accuracy was observed for fear and non-emotional samples, while anger proved to be the most difficult category to distinguish. Misclassifications tended to occur between emotions with similar facial action patterns, as revealed by SHAP interpretability analyses. For instance, anger was often confused with fear or sadness, reflecting its lack of a distinct classification pattern, whereas fear was frequently misclassified as surprise due to overlapping theoretical AUs [11]. The non-emotional category was particularly challenging, as it does not follow consistent patterns and tends to appear more randomly compared to other emotions. Notably, the SHAP analyses confirmed that most emotional categories exhibit discriminative patterns consistent with FACS theory, providing empirical support for the validity of the learned representations.

These findings are consistent with prior literature. For instance, Zhang et al. [24] reported high recognition accuracy for happiness, sadness, surprise, and neutral using reconstructed facial surfaces; however, their results were obtained from posed expressions and validated within a Unity-dependent reconstruction environment, limiting ecological validity. Similarly, research in immersive gaming contexts [25] achieved strong classification performance for comparable emotions, but relied on dual-camera recordings of participants imitating expressions, which were subsequently labeled from 2D imagery. In contrast, our approach uses only the Meta Quest Pro's facial landmark activations to classify spontaneously elicited emotions. Moreover, the use of Meta blendshapes provides a degree of interpretability that is difficult to achieve in camera-based or muscle-sensing setups, enabling us to identify which facial movements contribute most to prediction outcomes. These methodological differences likely explain the performance variations observed across studies.

While our findings support the viability of combining spontaneous affective responses with synthetic facial reconstruction, several limitations remain. Certain emotions, particularly fear and anger, proved difficult to distinguish, highlighting the need for more targeted elicitation strategies capable of producing clearer and more sustained affective responses. Future work will therefore focus on improving emotion elicitation through the integration of virtual humans within narrative and conversational VR scenarios, enabling more naturalistic and context-driven emotional engagement.

In addition, increasing annotation reliability constitutes an important next step. We plan to expand the number and diversity of external annotators and explore annotation protocols to enhance labeling consistency and reduce subjective bias.

Beyond these improvements, the primary direction for advancing this work lies in incorporating temporal modeling approaches capable of capturing the dynamic evolution of facial expressions over time. Ongoing research is exploring sequence-based architectures, including Transformer models, CNN–LSTM hybrids, and MiniRocket, with the goal of improving emotion discrimination by leveraging

temporal dependencies that are not captured by static representations. Together, these developments aim to strengthen the robustness, interpretability, and ecological validity of emotion recognition pipelines in immersive HMD environments.

Declaration on Generative AI

During the preparation of this manuscript, the authors used generative AI tools to improve language quality, including grammar and readability. The AI tools were used only for language polishing and not for generating scientific content, analyses, results, or discussions. The authors reviewed and edited the output and take full responsibility for the final content of the manuscript.

References

[1] V. Mäkelä, C. MacArthur, J. Lee, D. Harley, Integrating virtual reality head-mounted displays into higher education classrooms on a large scale, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI '25, Association for Computing Machinery, New York, NY, USA, 2025. URL: https://doi.org/10.1145/3706598.3713690. doi:10.1145/3706598.3713690.

[2] M. Slater, Implicit Learning Through Embodiment in Immersive Virtual Reality, Springer Singapore, Singapore, 2017, pp. 19–33. URL: https://doi.org/10.1007/978-981-10-5490-7_2. doi:10.1007/978-981-10-5490-7_2.

[3] F. Reer, L.-O. Wehden, R. Janzik, W. Y. Tang, T. Quandt, Virtual reality technology and game enjoyment: The contributions of natural mapping and need satisfaction, Computers in Human Behavior 132 (2022) 107242. URL: https://www.sciencedirect.com/science/article/pii/S0747563222000644. doi:https://doi.org/10.1016/j.chb.2022.107242.

[4] M. A. Cidota, S. G. Lukosch, P. Dezentje, P. J. Bank, H. K. Lukosch, R. M. Clifford, Serious gaming in augmented reality using hmds for assessment of upper extremity motor dysfunctions, i-com 15 (2016) 155–169. URL: https://doi.org/10.1515/icom-2016-0020. doi:doi:10.1515/icom-2016-0020.

[5] T. Luong, N. Martin, A. Raison, F. Argelaguet, J.-M. Diverrez, A. Lécuyer, Towards real-time recognition of users mental workload using integrated physiological sensors into a vr hmd, in: 2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), 2020, pp. 425–437. doi:10.1109/ISMAR50242.2020.00068.

[6] N. A. J. De Witte, F. Buelens, G. Debard, B. Bonroy, W. Standaert, F. Tarnogol, T. Van Daele, Handheld or head-mounted? an experimental comparison of the potential of augmented reality for animal phobia treatment using smartphone and hololens 2, Frontiers in Virtual Reality Volume 3 - 2022 (2022). URL: https://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2022.1066996. doi:10.3389/frvir.2022.1066996.

[7] M. Shomoye, R. Zhao, Automated emotion recognition of students in virtual reality classrooms, Computers Education: X Reality 5 (2024) 100082. URL: https://www.sciencedirect.com/science/article/pii/S2949678024000321. doi:https://doi.org/10.1016/j.cexr.2024.100082.

[8] M. Croissant, G. Schofield, C. McCall, Theories, methodologies, and effects of affect-adaptive games: A systematic review, Entertainment Computing 47 (2023) 100591. URL: https://www.sciencedirect.com/science/article/pii/S1875952123000460. doi:https://doi.org/10.1016/j.entcom.2023.100591.

[9] I. H. Bell, J. Nicholas, M. Alvarez-Jimenez, A. Thompson, L. Valmaggia, Virtual reality as a clinical tool in mental health research and practice, Dialogues Clin. Neurosci. 22 (2020) 169–177.

[10] J. Marín-Morales, J. L. Higuera-Trujillo, A. Greco, J. Guixeres, C. Llinares, E. P. Scilingo, M. Alcañiz, G. Valenza, Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors, Scientific Reports 8 (2018) 13657. URL: https://doi.org/10.1038/s41598-018-32063-4. doi:10.1038/s41598-018-32063-4.

[11] P. Ekman, W. V. Friesen, J. C. Hager, Facial Action Coding System: The Manual, Research Nexus, Salt Lake City, UT, 2002.

[12] J. Hamm, C. G. Kohler, R. C. Gur, R. Verma, Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders, J. Neurosci. Methods 200 (2011) 237–256.

[13] E. G. Krumhuber, L. I. Skora, H. C. H. Hill, K. Lander, The role of facial movements in emotion recognition, Nature Reviews Psychology 2 (2023) 283–296. URL: https://doi.org/10.1038/s44159-023-00172-1. doi:10.1038/s44159-023-00172-1.

[14] C. N. W. Geraets, S. Klein Tuente, B. P. Lestestuiver, M. van Beilen, S. A. Nijman, J. B. C. Marsman, W. Veling, Virtual reality facial emotion recognition in social environments: An eye-tracking study, Internet Interv. 25 (2021) 100432.

[15] Y.-Q. Cong, L. Yurdum, A. Fischer, D. Sauter, Testing the ingroup advantage in emotion perception from dynamic posed and spontaneous expressions, Journal of Nonverbal Behavior 49 (2025) 489–503. URL: https://doi.org/10.1007/s10919-025-00492-1. doi:10.1007/s10919-025-00492-1.

[16] M. Zloteanu, E. G. Krumhuber, D. C. Richardson, Acting surprised: Comparing perceptions of different dynamic deliberate expressions, Journal of Nonverbal Behavior 45 (2021) 169–185. URL: https://doi.org/10.1007/s10919-020-00349-9. doi:10.1007/s10919-020-00349-9.

[17] F. Cabitza, A. Campagner, M. Mattioli, The unbearable (technical) unreliability of automated facial emotion recognition, Big Data & Society 9 (2022) 20539517221129549. URL: https://doi.org/10.1177/20539517221129549. doi:10.1177/20539517221129549. arXiv:https://doi.org/10.1177/20539517221129549.

[18] M. Perusquía-Hernández, S. Ayabe-Kanamura, K. Suzuki, Human perception and biosignal-based identification of posed and spontaneous smiles, PLoS One 14 (2019) e0226328.

[19] A. M. S. Gonzalez-Acosta, M. Vargas-Treviño, P. Batres-Mendoza, E. I. Guerra-Hernandez, J. Gutierrez-Gutierrez, J. L. Cano-Perez, M. A. Solis-Arrazola, H. Rostro-Gonzalez, The first look: a biometric analysis of emotion recognition using key facial features, Frontiers in Computer Science Volume 7 - 2025 (2025). URL: https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2025.1554320. doi:10.3389/fcomp.2025.1554320.

[20] S. Li, W. Deng, Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition, IEEE Transactions on Image Processing 28 (2019) 356–370. doi:10.1109/TIP.2018.2868382.

[21] K. Kroenke, R. L. Spitzer, J. B. Williams, The PHQ-9: validity of a brief depression severity measure, J. Gen. Intern. Med. 16 (2001) 606–613.

[22] C. D. Spielberger, R. L. Gorsuch, R. Lushene, P. R. Vagg, G. A. Jacobs, State–trait anxiety inventory self evaluation questionnaire, form Y (STAI), 2003.

[23] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, P. A. Reuter-Lorenz, Emotional category data on images from the international affective picture system, Behavior Research Methods 37 (2005) 626–630. URL: https://doi.org/10.3758/BF03192732. doi:10.3758/BF03192732.

[24] Z. Zhang, J. M. Fort, L. Giménez Mateu, Facial expression recognition in virtual reality environments: challenges and opportunities, Frontiers in Psychology Volume 14 - 2023 (2023). URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2023.1280136. doi:10.3389/fpsyg.2023.1280136.

[25] T. Ortmann, Q. Wang, L. Putzar, EmojiHeroVR: A study on facial expression recognition under partial occlusion from Head-Mounted displays (2024). arXiv:2410.03331.