Affective and Pleasurable Design, Vol. 123, 2024, 161–172
https://doi.org/10.54941/ahfe1004687

Enhancing User Immersion in Virtual Reality by Integrating Collective Emotions Through Audio-Visual Analysis

Sam Corpataux, Marine Capallera, Omar A. Khaled, and Elena Mugellini

HumanTech Institute, HEIA-FR, Fribourg, Switzerland

ABSTRACT

In the rapidly evolving field of virtual reality (VR), enhancing user immersion is a key challenge. This study introduces an innovative approach integrating both audio and image analysis to detect and elicit collective emotion in 360° videos. Our holistic approach employs enriched datasets to train models, including face extraction and emotion classification based on spectrograms and audio features. Predictions from these models are merged to reflect crowd emotion valence, enriching VR environments with targeted visual, auditory, and haptic stimuli. User tests feedback show improvements in immersion, indicating the haptic dimension's potent impact in VR environments. This research highlights the potential of synthesizing audio-visual analyses to enhance crowd emotion detection, promising more sophisticated, emotionally affective VR systems.

Keywords: Crowd emotion analysis, Affective computing, Virtual reality, Multimodal interaction

INTRODUCTION

Virtual reality (VR) has evolved into an innovative technology that transforms how we interact with the digital world, enhancing immersion across entertainment, education, healthcare, and industry. VR immerses users in a digital universe, effectively blurring the lines between virtual and real worlds. This immersive experience is crafted through a combination of vivid visuals and diverse sensory feedback—visual, auditory, haptic, and occasionally olfactory—essential for delivering a compelling and cohesive experience.

Understanding user emotions in VR is key to enhancing immersion. Emotions are complex, but we do know that they are contagious (Gaines, 2021; Zhang et al., 2023); if a user is immersed in a VR experience where the surrounding crowd exhibits joy, the user is likely to feel happiness. Conversely, if the crowd displays panic, the user might experience fear. Thus, studying the emotions of a crowd can provide critical insights into individual emotional responses and significantly enhance user immersion. However, while existing research on crowd behavior often targets safety enhancements through

crowd-monitoring, such as (Jadhav et al., 2023; Khan et al., 2020) focusing on panic detection to prevent potential accidents, these do not typically seek to deeply understand the mood of a crowd. For studies that do focus on collective emotion, they typically focus on either audio or visual cues independently, neglecting the potential of a combined approach that utilizes both modalities to capture the full spectrum of emotional dynamics in VR.

This research addresses gaps in existing methodologies by introducing a novel hybrid approach to emotion detection in VR, leveraging both auditory and visual cues from crowd. This dual-input strategy enhances the precision of emotional detection, offering deeper insights into how emotions influence user experience in VR. The contribution lies in both the methodological integration and its application to VR, where understanding and responding to collective emotions can significantly transform user engagement and immersion.

RELATED WORK

Detecting Collective Emotion

In most research on detecting collective emotions using audio or visual data, the focus is primarily on detecting the valence of the emotion. Valence distinguishes between positive and negative emotions, such as joy and sadness, while arousal represents the intensity of the emotion (calm or excited). In this article, references to emotion will refer to valence unless explicitly stated otherwise. Valence values are generally divided into three categories: positive, neutral, and negative.

The work of (Franzoni et al., 2020) convert crowd noises into spectrograms. Initially, she used a CNN to classify these spectrograms, followed by other Deep Learning techniques, claiming to achieve excellent results (over 90% accuracy). However, these results are questioned (Faisal et al., 2021), who note that the dataset used is relatively small and lacks diversity, thus not ruling out the possibility that these results could be biased by overfitting. To address this issue, they created a new, larger dataset containing an additional 1297 seconds of labelled crowd recordings. Their approach also varies as, in addition to using spectrograms, they develop a second model based on raw acoustic features.

Most studies on detecting collective emotion are based on images. Many of these are related to the "Emotion Recognition in the Wild Challenge" (EmotiW) (Dhall et al., 2013). This academic event aims to gather researchers from around the world to address the problem of human emotion recognition. During the 2017 and 2018 EmotiW competitions, substantial progress was made in recognizing group emotions in natural settings. (Tan et al., 2017) developed a hybrid system using CNNs for both facial and scene recognition to optimize emotion detection from expressions and context. (Gupta et al., 2018) introduced an advanced attention mechanism to prioritize faces by their relevance to collective emotion, improving analysis accuracy. Another team, (Wang et al., 2018), implemented a cascading attention mechanism to

refine the importance of each face for a detailed global emotion representation. (Guo et al., 2020) integrated skeleton detection, combining posture data with facial analysis for a comprehensive understanding of group emotions.

Two studies propose novel approaches to understanding group emotions. (Zhang et al., 2017) focus on crowd movement with an emphasis on security applications rather than direct emotion detection. (Ghosh et al., 2022) do not directly address measuring group emotions in terms of valence but aim to infer group cohesion. While cohesion and emotion share similarities, the most intriguing aspect of their work is the use of the GAF 3.0 (Ghosh et al., 2022) dataset for assessing cohesion.

Improving VR Immersion

To enhance user immersion in VR, it is crucial to engage their senses and alter their perception of reality. While there is limited research on correlating taste, smell, or haptic stimuli with emotions, studies on visual and auditory stimuli exist. (Wilms and Oberfeld, 2018) showed that hue, saturation, and brightness affect emotions, with bright, warm hues enhancing positive emotions and darker colors inducing negative ones. (Västfjäll, 2012) demonstrated that pitch and volume influence emotions, with increased volume and higher frequencies enhancing positive emotions, and lowered volume and boosted bass highlighting negative ones. However, there is a notable gap in research on the impact of haptic stimuli, such as vibrations, on emotional responses. This work aims to address this gap by exploring the potential of haptic feedback to enhance VR user immersion.

METHODOLOGY

Data Preparation

Existing datasets were chosen and balanced for this study. The ELSCE dataset (Faisal et al., 2021) is used for audio data, with additional recordings to balance the underrepresented negative class. Data cleaning was performed, such as removing non-informative segments from recordings, to improve data quality. Despite these efforts, perfect balance was not achieved due to challenges in sourcing high-quality recordings of unhappy crowds, but the dataset was improved (see Table I).

The Python library librosa was used to extract 32 audio features, including zero crossing rate, energy, spectral entropy, spectral spread, MFCCs, and chroma. To address model overfitting, less influential features were discarded, retaining only the 11 most important features: energy, spectral spread, and nine MFCCs.

To train models based on images, we used the GAF 3.0 dataset (Ghosh et al., 2022; Wang et al., 2020), which contains over 17,000 labeled images. These images vary widely due to different group sizes, from 2 to over 100 individuals, and diverse gatherings (weddings, protests, funerals, and concerts). As the dataset is already well-balanced, no modifications were necessary. However, since our emotion classification model focuses on facial emotions, it was essential to extract faces from each crowd image. This step is

crucial because facial detection models are designed to process single cropped face images. Different models were tested to extract multiple faces from crowd images, with RetinaFace performing the best. However, its effectiveness needed limitations; for images with large crowd, it could detect dozens or even hundreds of faces. While promising, detecting so many faces is unnecessary for assessing overall crowd emotion. Since emotions are contagious, analyzing a subset of the crowd suffices for accurate mood assessment. Moreover, processing more faces increases computational time, slowing down the model. Various reduction techniques were implemented when the number of detected faces exceeded a threshold. First, a size-based filter retained the largest faces (e.g., top 20%), as larger, detailed faces are easier to classify. Second, a visibility filter excluded faces with obscured landmarks, retaining those with at least three visible landmarks. Finally, a filter based on RetinaFace's confidence index was applied. These techniques refined the model to detect only the most visible faces in a crowd (see Figure 1).

Table 1. Total duration of audio file per classes (in sec.) in the original and updated ELSCE.

Classes	Original ELSCE dataset	Updated ELSCE dataset
Positive	1772	1457
Negative	561	970
Neutral	1464	1464
Total	3797	3891

Audio Based Models' Architecture

This project explored two methods for classifying emotions from audio signals: direct feature extraction from audio files and converting audio into spectrograms for input into another classification model. To maximize performance, both approaches were combined, leveraging complementary information from raw audio features and spectrogram visualizations. The extracted features were used as inputs for two foundational models: raw audio features fed into an XGBoost model, while spectrograms trained a Random Forest Classifier.

Optimal hyperparameters for each model were determined through random and grid searches. For details on the parameters used for the XGBoost and Random Forest Classifier models, please contact the lead author, Sam Corpataux.

The outputs of these two models are merged as input to a meta-learner, implemented as a logistic regression. The main idea is that base models may produce different errors on different parts of the dataset. The meta-learner can learn these patterns to adjust its predictions accordingly (see Figure 2).

Image Based Models' Architecture

Scientific literature demonstrates multiple approaches to predicting collective emotions from images, typically using facial recognition, pose or skeleton

recognition, and entire image analysis. Facial recognition is the most common and well-documented method, which is why it was chosen for this project. Instead of developing a new model, we used pre-trained models from the "HSEmotion" library developed by HSE University researchers (researchers, n.d.). This library includes models (Savchenko, 2023) based on MobileNet and EfficientNet architectures, trained with the extensive VGGFace2 dataset. After testing the four highest-performing models, we selected the enet_b2_7 model for its superior performance.

Figure 1: Filtering detected faces.

This model is trained to recognize seven basic emotions: Angry, Disgust, Fear, Happiness, Neutral, Sadness, and Surprise. For this study, assessing emotion valence in three categories (positive, neutral, negative) is sufficient. To integrate predictions from both sound-based and image-based models, outputs are standardized by categorizing the seven emotions into three groups:

- Happiness*: **Positive***
- *N*eutral*: **Neutral***
- Angry, Disgust, Fear et Sadness*: **Negative.***

However, surprise is more complex to classify, as it can be either positive or negative. Simply assigning it a fixed label could bias results. Our solution uses the detected emotions of others in the group to classify surprise as either Positive or Negative. In large groups, people tend to express similar emotions simultaneously, so using the detected emotions around the individual can solve this issue. The classification model returns a probability vector for each detected face. These probabilities are collected for each face where the emotion is not surprise and sorted into positive or negative categories. We then count them and determine the majority class to rank surprise.

After categorizing all outputs from the image-based classification model into three categories, the model merges individual results to generate a single crowd image prediction. This process counts the emotions in each category and assigns the majority category to the crowd image. In tie situations, class probabilities determine the outcome. The final result is a single value for the crowd image, analogous to the model's audio-based output (see Figure 3).

Final Fusion Architecture

Each model type processes video segments independently: audio models analyze ambient sounds, while image-based models assess facial expressions. Results are categorized into three emotional states and combined on a graph with time on the horizontal axis and emotional valence on the vertical (see Figure 4). To capture intermediate states, we introduce Mid-Positive and Mid-Negative categories. Fusion of outputs is achieved manually using predefined decision rules refined through testing, ensuring accurate reflection of combined data (see "Fusion models' results" section). Automated fusion using a meta-learner was not implemented due to insufficient data and resource constraints. The manual approach is effective and less complex. Integrating multiple sensory data is crucial for VR, ensuring a seamless emotional narrative. By manually merging outputs from both models, we achieve a reliable and nuanced representation of the viewer's emotional experience.

MODELS RESULTS

Audio Based Models' Results

An XGBoost Classifier was used for the raw audio features model, chosen for its resilience against overfitting due to its regularized framework and gradient boosting technique. It achieved an average accuracy of 94.09% on the test set. A model using spectro-temporal features was also tested, analyzing the sound spectrum over time to capture essential temporal and frequency information for decoding emotional cues. This model achieved an average accuracy of 97.04% on the test set. Integrating these models into a meta-learner framework further enhanced performance, combining the strengths of individual models and minimizing weaknesses. The meta-learner achieved an impressive average accuracy of 98.71% on the test set.

Image Based Models' Results

The collective emotion classification model based on images is the Enet_b2_7 (researchers, n.d.), designed to recognize the seven basic emotions. For evaluation, the FER-2013 dataset (Zahara et al., 2020), was used. The model achieves an average accuracy of 56.15%. While this figure might seem modest, it is reasonable given the complexity of a seven-class classification problem. The model struggles more with negative emotions like anger, fear, and sadness, but performs relatively well in detecting joy. Despite these metrics, the potential for enhancement is significant when combined with results from the sound-based model.

Fusion Model Results

The fusion model combines predictions from both sound and image-based models, applying detailed techniques before deployment on 360° VR videos. Without standard datasets of annotated 360° crowd videos, quantitative assessment is challenging. To estimate performance, we sourced several suitable 360° crowd videos from the internet. These videos revealed that while a simple fusion of image and sound model results were satisfactory, it could be improved with additional post-processing rules.

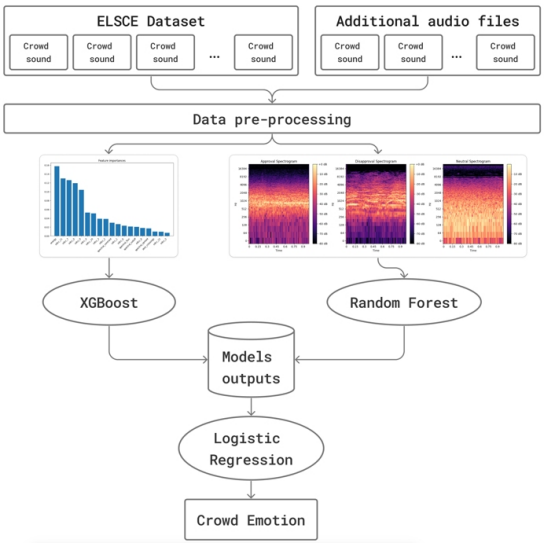

Figure 2: Audio based models' architecture.

Figure 3: Image based models' architecture.

The post-processing rules used to improve the fusion model's performance are:

i. Introduce a system to credit or weigh predictions for finer granularity, expanding the valence scale from 5 to 9 classes or levels.
ii. If the facial emotion classification model predicts 'Angry' while the sound-based model detects a 'Positive' emotion, consider the 'Angry' prediction a false positive and manually swap it with 'Happiness.'
iii. Use the predictions from the sound-based models to refine the classification of the 'Surprise' emotion.
iv. Attempt to detect and correct outliers by comparing the current prediction with the previous (i-1) and following (i + 1) fused predictions.

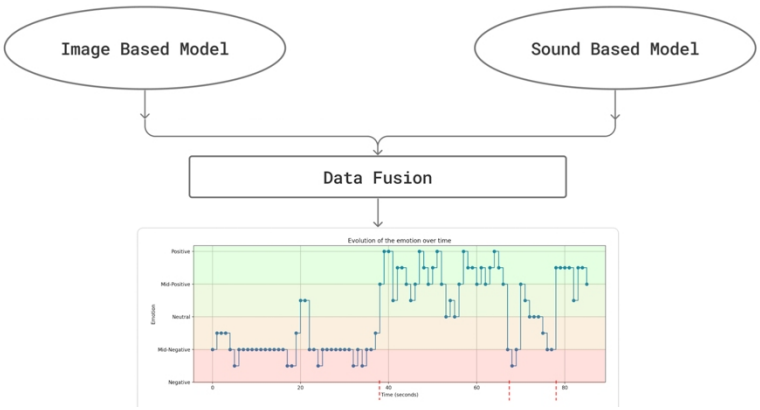

Figure 4: Fusion models' architecture.

All these rules are self-explanatory, except for Rule 2. This rule is applied because our model struggled to differentiate between faces screaming in joy and anger, often misclassifying both as anger. Screaming involves an open mouth, making the expression resemble anger more than joy. By using auditory input, we verify whether the scream is one of rage or happiness and reclassify if necessary.

EXPERIMENTS

Enhancing Immersion Techniques

To enhance VR user immersion, we focused on sight, sound, and haptic feedback. Visual filters were used to adjust hue, saturation, and brightness, with bright, warm hues for positive emotions and darker colors for negative emotions (Wilms and Oberfeld, 2018). Auditory adjustments involved modifying pitch and volume, increasing volume and higher frequencies for positive emotions, and lowering volume with boosted bass for negative emotions (Västfjäll, 2012).

These modifications were implemented using FFmpeg to segment, filter, and reassemble the video. Haptic feedback was provided through bHaptics vest to align with visual and auditory changes.

User Test Methodology

To evaluate the effectiveness of immersive enhancement techniques in virtual reality, we recruited 16 participants (8 males, 8 females) divided equally into a control group (CG) and an experimental group (EG). The CG experienced VR without additional stimuli, while the EG encountered the same VR scenarios but with added visual, auditory, and haptic stimuli. The participants, ranging in age from 22 to 35 ($M = 26{,}35$ years old, $SD = 3{,}16$), and declared varied levels of familiarity with VR.

The testing procedure was divided into four main steps:

1. Each participant began by completing the Immersive Tendencies Questionnaire (ITQ) (Witmer and Singer, 1998) to evaluate their predisposition to immersion.

2. They then engaged in the VR experience, with the EG receiving additional sensory stimuli tailored to the content of the VR.
3. Following the VR session, participants filled out the Igroup Presence Questionnaire (IPQ) (Schubert, 2003), which measures the sense of presence within the virtual environment.
4. The process concluded with a final interview to collect detailed feedback on their experiences and particularly their reactions to the sensory modifications.

This approach quantifies how visual, auditory, and haptic enhancements influence the user's immersion and sense of presence within a VR environment, providing a robust dataset for analyzing the effectiveness of these immersive techniques.

User Test Results

User test results confirm the effectiveness of sensory stimuli in enhancing virtual reality immersion. Both EG and CG began with similar scores on the Immersive Tendencies Questionnaire (ITQ), 80.75 and 81.13 out of 126, respectively, providing a consistent baseline. The EG, exposed to additional sensory enhancements, scored 45.88 out of 63 on the Igroup Presence Questionnaire (IPQ), compared to the control's 40.38, an 8.6% improvement. Feedback from interviews highlighted the impact of haptic feedback in amplifying feelings of joy and euphoria, though visual and auditory enhancements were less perceptible, with some participants mistaking visual filters for technical errors. A T-Test yielded a p-value of 0.000369, indicating statistically significant differences. These insights suggest that while haptic feedback significantly boosts immersion, auditory and visual enhancements need refinement for better perceptibility and impact.

DISCUSSION

Our research into enhancing user immersion in virtual reality by integrating collective emotions through audio-visual analysis has shown promising results and identified key areas for future development. By combining feedback from various sources, we gained a better understanding of emotional dynamics in VR settings. However, improvements are needed, particularly in our image-based models. Currently, these models focus mainly on facial expressions to assess emotions. Incorporating body pose analysis could provide deeper insights into individuals' emotional states within a crowd. Also, integrating object or scene detection models could enhance context understanding, refining emotion classification. For instance, recognizing environments or objects like protest banners or celebration decorations could help deduce the crowd's mood and event nature accurately.

Additionally, automating the fusion of audio and visual model outputs through a meta-model could enhance efficiency and accuracy compared to manual rules. Expanding our audio dataset would further improve model robustness and reliability, especially in diverse scenarios. These enhancements

could significantly advance VR technologies, creating more immersive environments capable of dynamically reflecting human emotions. Such advancements could revolutionize fields from entertainment to therapy. In summary, while our study has established a foundation for integrating collective emotions into VR, exploring these improvements could greatly enhance our systems' functionality and applicability, fostering emotionally aware virtual environments.

CONCLUSION

In this project, we introduced an innovative approach for group-level emotion recognition in virtual reality environment. By integrating audio-visual analysis, we achieved enhanced user immersion by accurately capturing collective emotions. Our models, based on XGBoost and Random Forest for audio and EfficientNet for images, demonstrated high precision in emotion classification. Through manual fusion of audio and visual results, we developed a robust system for recognizing collective emotions in VR. Furthermore, user tests revealed a notable improvement in immersion for participants exposed to sensory enhancements, particularly through haptic feedback. While our approach is promising, future work should focus on refining image models, automating fusion processes, and expanding audio datasets for improved reliability and efficiency. Overall, our work lays the groundwork for emotionally responsive virtual environments with applications across entertainment, education, and beyond.

ACKNOWLEDGMENT

This project was funded by HEIA-FR. Special thanks to Dr. Marine Capallera for her secondary authorship, review, and feedback. We also acknowledge the support of Prof. Elena Mugellini and Prof. Omar A. Khaled from the HumanTech institute.

REFERENCES

Atick Faisal, M. A., Ahmed, M. U., Rahman Ahad, M. A., 2021. ESLCE: A Dataset of Emotional Sounds from Large Crowd Events. 2021 Joint 10th International Conference on Informatics, Electronics & Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision & Pattern Recognition (icIVPR) 1–7. https://doi.org/10.1109/ICIEVicIVPR52578.2021.9564179

Dhall, A., Goecke, R., Joshi, J., Wagner, M., Gedeon, T., 2013. Emotion Recognition In The Wild Challenge (EmotiW) challenge and workshop summary: 2013 15th ACM International Conference on Multimodal Interaction, ICMI 2013. ICMI 2013 - Proceedings of the 2013 ACM International Conference on Multimodal Interaction 371–372. https://doi.org/10.1145/2522848.2531749

Franzoni, V., Biondi, G., Milani, A., n.d. Crowd emotional sounds: spectrogram-based analysis using convolutional neural networks.

Ghosh, S., Dhall, A., Sebe, N., Gedeon, T., 2022. Automatic Prediction of Group Cohesiveness in Images. IEEE Transactions on Affective Computing 13, 1677–1690. https://doi.org/10.1109/TAFFC.2020.3026095

Guo, X., Polania, L., Zhu, B., Boncelet, C., Barner, K., 2020. Graph Neural Networks for Image Understanding Based on Multiple Cues: Group Emotion Recognition and Event Recognition as Use Cases. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV).

Gupta, A., Agrawal, D., Chauhan, H., Dolz, J., Pedersoli, M., 2018. An Attention Model for Group-Level Emotion Recognition, in: Proceedings of the 20th ACM International Conference on Multimodal Interaction, ICMI '18. Association for Computing Machinery, New York, NY, USA, pp. 611–615. https://doi.org/10.1145/3242969.3264985

Jadhav, C., Ramteke, R., Somkunwar, R. K., 2023. Smart Crowd Monitoring and Suspicious Behavior Detection Using Deep Learning. | Revue d'Intelligence Artificielle | EBSCOhost [WWW Document]. https://doi.org/10.18280/ria.370416

Khan, A., Ali Shah, J., Kadir, K., Albattah, W., Khan, F., 2020. Crowd Monitoring and Localization Using Deep Convolutional Neural Network: A Review. Applied Sciences 10, 4781. https://doi.org/10.3390/app10144781

Ph. D, J. G., 2021. What Is Emotional Contagion Theory? (Definition & Examples) [WWW Document]. PositivePsychology.com. URL https://positivepsychology.com/emotional-contagion/ (accessed 4.26.24).

Regenbrecht, H., Schubert, T., 2002. Real and illusory interactions enhance presence in virtual environments. Presence: Teleoperators & Virtual Environments 11, 425–434.

Researchers, H. U., n.d. HSEmotion (High-Speed face Emotion recognition) library.

Savchenko, A. V., 2023. MT-EmotiEffNet for Multi-task Human Affective Behavior Analysis and Learning from Synthetic Data, in: Proceedings of the European Conference on Computer Vision (ECCV 2022) Workshops. Springer, pp. 45–59.

Schubert, T. W., 2003. The sense of presence in virtual environments: A three-component scale measuring spatial presence, involvement, and realness. Z. für Medienpsychologie 15, 69–71.

Tan, L., Zhang, K., Wang, K., Zeng, X., Peng, X., Qiao, Y., 2017. Group Emotion Recognition with Individual Facial Emotion CNNs and Global Image Based CNNs, in: Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI '17. Association for Computing Machinery, New York, NY, USA, pp. 549–552. https://doi.org/10.1145/3136755.3143008

Västfjäll, D., 2012. Emotional reactions to sounds without meaning. Psychology 3, 606.

Wang, K., Zeng, X., Yang, J., Meng, D., Zhang, K., Peng, X., Qiao, Y., 2018. Cascade Attention Networks For Group Emotion Recognition with Face, Body and Image Cues, in: Proceedings of the 20th ACM International Conference on Multimodal Interaction, ICMI '18. Association for Computing Machinery, New York, NY, USA, pp. 640–645. https://doi.org/10.1145/3242969.3264991

Wang, Y., Wu, J., Huang, J., Hattori, G., Takishima, Y., Wada, S., Kimura, R., Chen, J., Kurihara, S., 2020. LDNN: Linguistic Knowledge Injectable Deep Neural Network for Group Cohesiveness Understanding, in: ICMI '20. Association for Computing Machinery, New York, NY, USA.

Wilms, L., Oberfeld, D., 2018. Color and emotion: effects of hue, saturation, and brightness. Psychological research 82, 896–914.

Witmer, B. G., Singer, M. J., 1998. Measuring presence in virtual environments: A presence questionnaire. Presence 7, 225–240.

Zahara, L., Musa, P., Prasetyo Wibowo, E., Karim, I., Bahri Musa, S., 2020. The Facial Emotion Recognition (FER-2013) Dataset for Prediction System of Micro-Expressions Face Using the Convolutional Neural Network (CNN) Algorithm

based Raspberry Pi, in: 2020 Fifth International Conference on Informatics and Computing (ICIC). Presented at the 2020 Fifth International Conference on Informatics and Computing (ICIC), pp. 1–9. https://doi.org/10.1109/ICIC50835.2020.9288560

Zhang, L., Chen, Y., Wei, Y., Leng, J., Kong, C., Hu, P., 2023. Kick Cat Effect: Social Context Shapes the Form and Extent of Emotional Contagion. Behavioral Sciences 13, 531. https://doi.org/10.3390/bs13070531

Zhang, Y., Qin, L., Ji, R., Zhao, S., Huang, Q., 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, 635–648. https://doi.org/10.1109/TCSVT.2016.2593609