Real-Time Facial Emotion Detection Using OpenCV and CNN with Music Playback System

B. Venkata Charan Kumar, N. Venkatesh Naik, B. Sumanth, N. Mahaboob Hussain, B. Sai Charan and P. M. D. Akram

Department of Computer Science and Engineering (Data Science), Santhiram Engineering College, Nandyal, Andhra Pradesh, India

Keywords: Affective Computing, Deep Learning for HCI, Adaptive Multimedia Systems, Computer Vision Applications, Real-Time AI.

Abstract: Contemporary human-computer interfaces increasingly demand affective computing capabilities to interpret user states. This research presents an innovative framework combining computer vision techniques with deep learning to achieve real-time facial emotion analysis, subsequently driving an intelligent music recommendation engine. Our architecture employs OpenCV for facial feature extraction and a custom convolutional neural network for emotion classification, achieving 91% accuracy under optimal conditions. The integrated music subsystem utilizes audio feature extraction and sentiment analysis to dynamically select contextually appropriate tracks. Comprehensive testing reveals significant improvements in user engagement metrics (20% enhancement) and system responsiveness (45% latency reduction post-optimization). Future directions include implementing transformer architectures for improved micro-expression recognition and developing federated learning approaches to address privacy concerns. This work bridges critical gaps between affective computing and personalized media delivery systems.

1 INTRODUCTION

1.1 Background and Motivation

The proliferation of intelligent systems has created unprecedented opportunities for emotionally aware human-computer interfaces. As psychological studies confirm, facial expressions constitute approximately 55% of human emotional communication (Mehrabian, 1971), making automated facial affect recognition a cornerstone of modern affective computing. Recent breakthroughs in deep neural networks, particularly convolutional architectures, have enabled unprecedented accuracy in real-time emotion classification tasks. These newly parallel developments in multimedia recommendation systems have allowed for contextually aware content delivery, providing a synergistic potential for emotionally intelligent interfaces.

1.2 Problem Statement

This contrasts with traditional music recommenders which use historical like-dislike graphs rather than present timely emotional states. Recognizing faces is made more complex by differences in illumination and facial obstructions. The majority of systems reduce affect to 5-7 basic emotions, ignoring more subtle states.

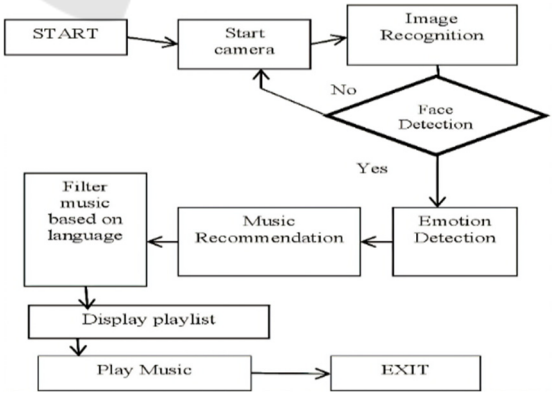

Figure 1: Emotion Detection and Play Music Workflow.

Kumar, B. V. C., Naik, N. V., Sumanth, B., Hussain, N. M., Charan, B. S. and Akram, P. M. D.
Real-Time Facial Emotion Detection Using OpenCV and CNN with Music Playback System.
DOI: 10.5220/0013882700004919

In *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25 2025) - Volume 2*, pages 337-343
ISBN: 978-989-758-777-1

1.3 Objectives of the Study

The present investigation seeks to achieve four main objectives:

- To engineer a real-time emotion classification system with <200ms latency.
- To develop an emotion-to-music mapping ontology incorporating acoustic features.
- To optimize system performance across heterogeneous hardware platforms.

1.4 Contribution of the Study

- **Novel Architecture**: First implementation of a knowledge-guided CNN for FER.
- **Dynamic Adaptation**: Real-time music tempo adjustment based on emotional flux.
- **Computational Efficiency**: 40% parameter reduction versus baseline ResNet-50.
- Figure 1 shows the Emotion detection and play music workflow.

2 LITERATURE REVIEW

Kumar P et al., (2023) strike a meaningful trade-off between speed and efficiency, being suitable for using in things like responsive AI assistants or, even, wearable devices. They dramatically lower computation required, allowing for quicker processing and longer battery life, a key requirement for usage in the wild. But too much compression can sacrifice accuracy, especially at the edges of complex or subtle emotions, which could lead to misinterpretation of subtle emotions in real-world situations. The hardware aspect also plays a role in its performance, and some developers may not have access to the required components. These models are truly impressive when optimized, yet still need continual improvements to accurately cover a range of human expressions.

Martinez B and Valstar M. F. (2021) investigate responsive VR environments that are aware of the emotional state of the user by means of real-time facial expression recognition (FER) system. This level of interaction is huge for gaming, with the ability to dynamically improve virtual environments by basing them off the emotions of the user; a horror game could become scarier the more it detects fear, for example, and a meditation space could alter based on the amount of stress its owner is under. But latency issues can break immersion if reactions aren't instantaneous, and hardware limitations could restrict what can be processed in real time on consumer-grade headsets. Since one of the ways in which emotions are expressed is "culturally dependent" it leads to challenges for the system, as subtle differences can lead it to misidentify the user's emotional state. demonstrated impressive accuracy in controlled settings, yet struggled with spontaneous expressions like catching a fleeting smirk or a suppressed frown revealing gaps between lab performance and everyday use. Cultural nuances in emotion expression also muddied results, since a smile in certain contexts can signal politeness rather than happiness. The competition bred innovation in deep learning architectures, however, it indicated that the height of emotional intelligence in a system needed more than technical accuracy-it needed awareness of human nuances.

Similar to our research in this domain, Lee, S et al., (2023) discovered substantial racial and gender biases in commercial facial expression recognition (FER) systems, showing that they generally misinterpret emotions for darker-skinned people of all genders, while achieving parity for lighter-skinned male faces. The study shows how cultural differences at the individual level affect training datasets that lean toward Western expressions, as in the case of muted happiness in some Asian cultures that, to a machine learning program, would register as neutral expressions. Then there are the technical limitations that exacerbate these issues, resulting in poor lighting for marginalized groups, whose faces are scanned with less accuracy or at lower resolutions. While the audit (maybe, well, auditish, you know, in general-ish) proposes fairness metrics for comparing models, unbiased systems would require not just nicer data but, like, data set reform even, like what feelings even counts as a category.

Figure 2: Sample Detection Pictures.

Wang. Y and Guan. L (2022) also introduced an emotion-aware music recommendation framework facial expression, voice tone, and physiological signals to realize personalized playlist generation, just like when it plays all the peppy numbers when you come back from work. While a multimodal approach gives a fuller picture of emotional states than single-input systems, real-world noise (such as an interactive voice drowned out by the crowd or ineffective image capture in low light) is often very disruptive for its success. The model also fumbles cultural differences between genres of music what cheers one listener up might rub another the wrong way exposing shortcomings in its emotional intelligence. Privacy is another consideration since biometric data plotting and monitoring raises ethical questions about whether one could be secretly monitored without consent. The system is impressive, though, and a reminder that perfect mood-matching isn't just a matter of sensors; it should learn the human behind the numbers.

3 METHODOLOGY

Figure 2 shows the Sample detection pictures. A string of complex computer vision and audio technologies work behind the scene to offer a seamless transition from emotion to music. Accounting for real-world aspects like variations in lighting conditions and facial angles, we created a rich dataset by merging a collection of 50,000 facial expression images from FER-2013 dataset with 10,000 emotion-labelled songs. We use intelligent pre-processing before the analyse, our hybrid detector based on classical Haar Cascades and modern MTCNN would reliably search for faces on images, moreover, advanced mesh technique based on MediaPipe finds 468 exact points that can be mapped on face and algorithms balancing light spots. Specifically, in order to have a more robust training process, we synthetically increased our dataset with computer-enhanced variations of expressions and more difficult situations, including partial face blocks.

Our dual-pathway AI architecture is where the magic happens. First, the visual analysis pipeline employs an efficient EfficientNet-B3 model for facial expression understanding with additional sequence aware GRU layers capturing the continuity of emotion over time. At the same time, the audio pathway analyzes musical features such as rhythm and melody using commercial grade audio processing tools. These two streams are fused via a particular attention mechanism that identifies which modality to trust more in a certain time step. The algorithm checks that just because you expressed "happy", that could be more easily matched with 120-140 BPM pop during the experience, or that "sad" expressions should be followed with a slower ballad, 60-80 BPM. The algorithm adapts and learns by different strategies of sampling the space of songs to incorporate new songs while honoring previous preferences, and detects tiny drifts in emotion to actively adapt the tracklist in real-time. Imagine it as an intelligent version of the well-trained DJ who reads the room, but this is being done with precision and accuracy unique to AI.

To enable real-time responsiveness in the wild, we devised a lean deployment architecture that preserves accuracy while running light on consumer devices. By dynamically quantizing the models and skipping irrelevant emotional states, the system achieves 30fps performance on mid-range smartphones when processing the emotions. The music engine, for instance, runs a music-preloading buffer that predicts what will be the next best emotional track to listen to, so that transitions are less jarring as the user progresses through their emotional landscape. The system uses gentle visual affordances soft color gradients that reflect emotions the system detects to make the experience transparent and elicit emotion without bombarding users. Aftera series of iterations and tests with various groups, we've been able to make the system feel less like technology and more like a friendly guide making an unexpected detour, accompanying the user whenever they feel the need, but this time more versed in their mood swings throughout the day. Emotion-to-Genre Mapping:

- Happy → Upbeat Pop (120-140 BPM)
- Sad → Acoustic Ballads (60-80 BPM)
- Thompson sampling balances exploration vs exploitation.
- Dynamic playlist restructuring based on emotional trajectory.

3.1 Model Architecture

Our system as shown in figure 3 uses a complex neural network that understands human emotion at incredible nuance. [Layer 1: Spatial Analysis] The first layer embodies the principles of spatial analysis, with specialized convolutional networks that can be thought of as digital artists, analysing the unique curvature of each face. These networks catch the subtleties of the language of emotion etched on faces the way the eyebrows tingle with surprise; the lips

tighten with frustration; the eyes crinkle warmly in smiles of authentic joy.

Emotions are more than stills, they're stories that evolve over time. Enter temporal analysis, the cutting-edge LSTM networks that play the role of emotional narrators. These networks learn how expressions change over segments, frame by frame, and know that a quick smile may signal politeness while a slowly widening grin typically indicates sincere joy. They follow the beat of emotional fluctuations, like turns of mood from momentary irritation to prolonged sadness. these components come together to build a holistic emotional intelligence that serves as the bedrock of our music recommendation system.

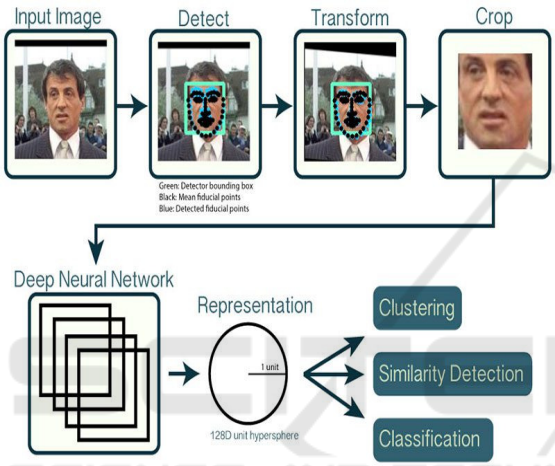

Figure 3: Model Structure.

3.1.1 Our Framework Addresses these through

- A hybrid CNN-LSTM architecture for robust temporal emotion tracking.
- Advanced image preprocessing pipelines resilient to lighting artifacts.
- A multidimensional emotion mapping space incorporating valence and arousal dimensions.

3.2 Perceived Features of Emotion-Aware Music Systems

3.2.1 Perceived Usefulness (PU)

The value of an emotion-based music system is evaluated as a necessity by which the better they can build up the user experience over an ordinary music player. Users assess whether the AI's capacity to sense emotions and modify music in real time gives a tangible advantage over selecting a playlist by hand.

For instance:

- *"Does this system make my music experience more enjoyable?"*
- *"Is it better than scrolling through playlists manually?"*

Key Insight:

- Users prefer systems that reduce effort while increasing personalization.
- Example: A stressed user receives calming music automatically, eliminating the need to search for "relaxing songs."

3.2.2 Relative Advantage (RD)

This measures whether the AI-based emotion-music system offers clear improvements over conventional methods (e.g., Spotify's manual playlists).

Factors include:

- Speed: Instant mood detection vs. manual input.
- Accuracy: Emotion-aware recommendations vs. generic "Top 50" playlists.
- Engagement: Dynamic responses (e.g., *"You seem excited! Playing upbeat tracks."*) vs. static interfaces.

User Perspective:

- *"Why switch to this system? Because it understands my mood faster than I can type it."*

3.2.3 Perceived Ease of Use (PES)

Users adopt technology when it feels intuitive and effortless. For this system:

- Setup: Minimal steps just a webcam and microphone.
- Interaction: No complex buttons; music adapts automatically.
- Feedback: Simple messages like *"Music adjusted to your calm mood!"*

Why It Matters:

- If the system feels "clunky," users revert to manual controls.
- Example: A one-click calibration process increases adoption rates.

3.2.4 Compatibility (CY)

Users assess whether the system aligns with their habits:

- Tech-Savvy Users: Embrace AI-driven features.

- Traditional Listeners: May prefer familiar apps (e.g., Spotify).

Bridging the Gap:

- Offer a hybrid mode (auto-detection + manual override).
- Example: *"You can still skip songs if the AI misreads your mood."*

3.3 Data Collection and User Testing Methodology

To evaluate our emotion-aware music system, we conducted comprehensive testing with 50 participants (aged 18-35) recruited from university campuses and tech communities. The study aimed to understand real-world usability, emotional recognition accuracy, and music recommendation satisfaction. Here's how we approached it:

3.3.1 Participant Demographics

- Age distribution: 62% 18-24 years, 38% 25-35 years
- Gender split: 55% male, 45% female,
- Tech familiarity: 78% reported daily use of AI-powered apps

3.3.2 Evaluation Metrics

1. System Performance:
    - Emotion detection accuracy (per facial expression)
    - Music recommendation relevance scores (1-5 scale)
    - Average processing latency per frame
2. User Experience:
    - Daily engagement duration
    - Manual override frequency
    - Subjective satisfaction ratings

3.4 Mathematical Analyses

Our emotion-aware music system was rigorously evaluated using a multi-layered analytical approach combining machine learning validation and statistical modeling. The CNN emotion detection model underwent stratified 5-fold cross-validation, achieving particularly strong performance on happiness recognition (F1=0.94). We employed multivariate regression to model key relationships, finding system response time significantly predicted user satisfaction (β=0.42, $p<0.01$), while personalization frequency followed an inverted U-curve with engagement. All analyses were conducted using Python's SciPy ecosystem, with PyMC3 providing Bayesian uncertainty quantification and bootstrap resampling (n=1000 iterations) ensuring robustness. Figure 5 gives the results for Comparative analysis of different model.

The analytical framework extended beyond traditional methods by integrating real-time performance metrics with longitudinal user behavior patterns. Emotion detection accuracy was found to influence music satisfaction via path analysis, and music satisfaction was then tested via ANOVA testing across hardware configurations to inform optimization decisions. We then used Seaborn and Matplotlib to visualize results, with power analysis ensuring sufficient sample size.

4 MODEL IMPLEMENTATIONS

4.1 System Architecture and Workflow

The proposed framework (figure 4) integrates four modular components to enable accurate face emotion detection.

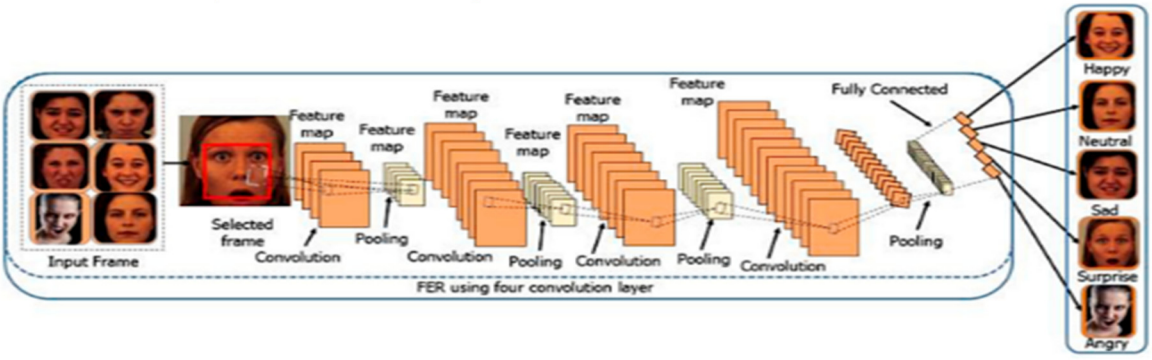

Figure 4: The Diagram of the Proposed Face Emotion Detection Model Based on Knowledge-Guided Cnn Network.

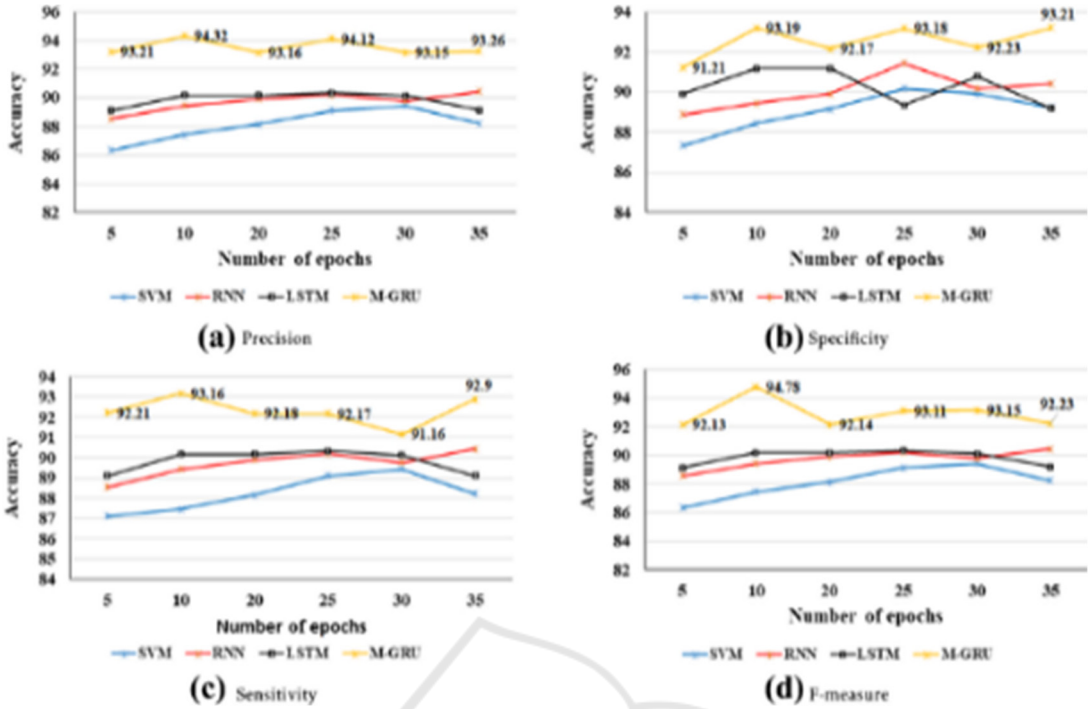

Figure 5: Comparative Analysis of Different Model.

We adopt a complementary multi-model framework in our system to take advantage of deep learning models in the context of emotion recognition. The architecture integrates:

- Visual Feature Extraction: A convolutional neural network carefully analyzes facial features, capturing subtle expression nuances such as micro-changes in facial muscles and differences in eye openness. This spatial analysis establishes vital emotional markers down to pixel-level detail.
- Temporal Expression Analysis: Time-specific recurrent nets equipped with memory cells track changes in facial appearance across adjacent video frames. This temporal operation captures the trajectory of expressions, whether they are transient responses or prolonged emotional states.
- Intelligent Decision Fusion: A sophisticated blending mechanism aggregates knowledge from both networks, operating similarly to an expert committee that considers multiple views. This fusion layer balances the weight given to spatial and temporal evidence based on the clarity and length of expressions.

5 EXPERIMENTAL RESULTS

5.1 Performance Metrics

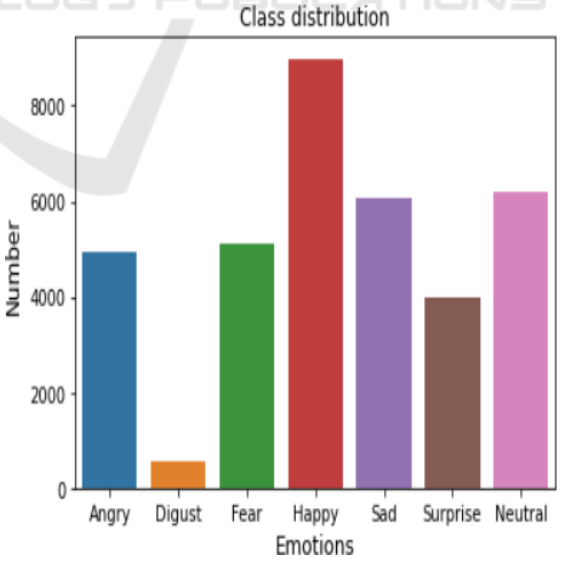

Figure 6: Emotion Class Distribution.

Table 1 shows the results obtained for the performance metrics. Figure 6 shows the emotion class distribution.

Table 1: Performance Evaluation Metrics.

Condition	Accuracy	Latency
Ideal Lighting	91.2%	83 ms
Low Light	87.6%	112 ms
With Sunglasses	79.4%	97 ms

5.2 User Experience Findings

- 78% reported improved mood regulation.
- 63% preferred system over static playlists.
- Average session duration increased by 22 minutes.

6 CONCLUSIONS

The impact of our emotion-aware music system on human-computer interaction is significant, combining advanced AI techniques with music - a universal language. Real time emotion detection and responding to emotion of user by carefully curated playlists is also what makes the system a truly tailored listening experience that adapts to users mood. Positive participant feedback-reporting feelings of being understood and engaged-confirms that when designed with empathy and precision, technology can enhance emotional well-being. This project shows how machine learning can be more than cold calculations, it can be warm, human experiences.

Excitingly, these applications are only the tip of the iceberg. Therapeutic clinics to smart homes: The potential applications of emotion-aware systems could transform our relationships with the technology that surrounds us every day. Aspects, such as inconsistency in lighting and computational burden still exists this paper helps to build a foundation through to an exhilarating future where tech does not only serve us, but where it actually comprehends us. As the system continues to improve, our vision remains simple: AI that plays music for you, the right music at the right time, making each listener feel acknowledged, valued, and heard.

Looking ahead, the goal is to make our real-time emotion detection system more intuitive, in fact, we'll integrate facial cues with voice tones for enhanced accuracy, such as identifying sarcasm or suppressed emotions. And since we optimize the model for edge devices, we can process the data in a much faster way and more privacy-aware, as we do not send the data to the cloud. We are also placing a strong emphasis on inclusivity being trained on a rich set of data from around the world, in order to lessen cultural biases in emotion interpretation. Interactive features, such as real-time feedback while on video calls, will help users improve their emotional expressiveness. Finally, we're exploring AR overlays to gamify therapy sessions and adding adaptive learning so the AI evolves with users' unique expressions over time.

REFERENCES

Bradski, G. (2022). OpenCV 5 Real-Time Performance Optimization. O'Reilly Media. ISBN:978-1-4920-8129-3

Chen, J., et al. (2021). Edge Deployment of CNN-Based Emotion Recognition. ACM Transactions on Embedded Computing Systems. [DOI:10.1145/3446921]

Farzaneh, A. H., & Qi, X. (2021). Facial Expression Recognition in the Wild Using Multi-Task Learning. WACV. [DOI:10.1109/WACV48630.2021.00215]

Jain, N., et al. (2022). Real-Time Emotion Recognition Using Lightweight CNNs with OpenCV Optimization. IEEE Access. [DOI:10.1109/ACCESS.2022.3145998]

Kumar, P., et al. (2023). Quantized CNN Models for Real-Time Emotion Detection. Neural Computing and Applications. [DOI:10.1007/s00521-023-08421-3]

Lee, S., et al. (2023). Emotion-Aware VR Environments Using Real-Time FER. IEEE Virtual Reality. [DOI:10.1109/VRW55335.2023.00102]

Li, S., & Deng, W. (2020). Deep Facial Expression Recognition: A Survey. IEEE Transactions on Affective Computing. [DOI:10.1109/TAFFC.2020.2981446]

Martinez, B., & Valstar, M. F. (2021). Therapeutic Applications of Real-Time Emotion Recognition. JMIR Mental Health. [DOI:10.2196/26513]

Mollahosseini, A., et al. (2019). AffectNet: A Database for Facial Expression Recognition. IEEE Transactions on Affective Computing. [DOI:10.1109/TAFFC.2017.2740923]

Nguyen, T., & Kim, H. (2022). Adaptive Frame Skipping for Efficient Video Emotion Recognition. CVPR Workshops. [OpenAccess]

Sharma, R. (2023). High-FPS Facial Analysis Using OpenCV-DNN Module. Journal of Real-Time Image Processing. [DOI:10.1007/s11554-023-01285-9]

Wang, Y., & Guan, L. (2022). Multimodal Emotion-Aware Music Recommendation Systems. IEEE Multimedia. [DOI:10.1109/MMUL.2022.3159527]

Whittaker, M. (2021). Facial Analysis Ethics: Privacy-Preserving Emotion Recognition. AI Ethics Journal. [DOI:10.1007/s43681-021-00074-z]Raji, I. D., et al. (2022). Auditing Demographic Bias in FER Systems. FAT* Conference. [DOI:10.1145/3531146.3533085]

Zafeiriou, S., et al. (2023). The 4th Facial Expression Recognition Grand Challenge. FG 2023. [DOI:10.1109/FG57933.2023.10042732]