International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 22s, 2025
https://theaspd.com/index.php

Enhancing Affective Modelling Through Preprocessing: A Comparative Review for Valence and Arousal Estimation

G. Divya[1], P. Hemalatha[2], Soumya Jolad[3], Neeru Yadav[4]
[1]Department of Computer science and Engineering East Point College of Engineering and Technology, Approved by AICTE New Delhi, Affiliated to VTU Belagavi, Bengaluru, Karnataka, India. divyavemareddy97@gmail.com
[2]Department of Computer Science and Engineering East Point College of Engineering and Technology, Approved by AICTE
New Delhi, Affiliated to VTU Belagavi, Bengaluru, Karnataka, India. hemaece40@gmail.com
[3]Department of Computer Science and Engineering, East Point College of Engineering andTechnology, Approved by AICTE New Delhi, Affiliated to VTU Belagavi, Bengaluru, Karnataka, India. Soumyajolad4@gmail.com
[4]Department of Computer Science and Engineering, East Point College of Engineering andTechnology, Approved by AICTE New Delhi, Affiliated to VTU Belagavi, Bengaluru, Karnataka, India. raoneeru92@gmail.com

Abstract: *Facial emotion recognition, especially the estimation of valence and arousal, plays a critical role in affective computing and human-computer interaction. The reliability of these predictions is significantly influenced by the preprocessing techniques used on facial images. This paper presents a comparative review of conventional and advanced preprocessing approaches, ranging from traditional image normalization to deep learning-driven methods, including data augmentation, multimodal fusion, and spatiotemporal encoding. Through critical analysis and referencing of recent literature, we highlight how preprocessing impacts model performance and suggest best practices for enhancing affective modeling pipelines.*

Keywords: *Facial Emotion Recognition, Deep Learning Preprocessing Techniques, Multimodal Fusion, Temporal-Aware Features Introduction.*

1. INTRODUCTION

Affective computing seeks to interpret human emotions, with facial image analysis serving as a central modality. Among emotion metrics, valence (positivity/negativity) and arousal (intensity) are widely used. Preprocessing of facial images is foundational to improving the accuracy of automated emotion recognition systems. This review compares major preprocessing techniques and explores their contributions to model performance.

2. LITERATURE REVIEW

A. Traditional Preprocessing Techniques

Traditional preprocessing forms the initial and essential stage in facial emotion recognition systems. These techniques are typically lightweight yet effective in improving data quality, reducing variability, and enhancing the performance of downstream feature extraction and model training processes.

a. Face Detection and Alignment

Accurate localization of the face within an image is a prerequisite for consistent and reliable feature extraction. Face detection techniques, such as those based on Haar cascades, offer a fast and computationally efficient method to identify facial regions. More recently, deep learning-based detectors have shown improved robustness under varied lighting and occlusion conditions.

Once the face is detected, alignment procedures are applied to standardize the position of facial landmarks such as the eyes, nose, and mouth. This alignment reduces inter-sample variance due to head pose or orientation and ensures that emotion-relevant regions are consistently framed. As highlighted by Dalal et al. (2023) and Meena and Velmurugan (n.d.), proper face alignment enhances the stability and accuracy of feature extraction, especially in systems that rely on convolutional filters sensitive to spatial location.

b. Grayscale Conversion and Normalization

Converting colour images to grayscale is a common preprocessing step that reduces input dimensionality while preserving essential structural information. This is particularly useful in facial emotion tasks, where color may not significantly contribute to emotional content but adds computational overhead.

Normalization complements grayscale conversion by adjusting the pixel intensity values to a standard range, typically between 0 and 1 or standardized to zero mean and unit variance. This process facilitates faster convergence during model training and ensures consistency across different samples. These practices are widely adopted and discussed in works such as Meena and Velmurugan (n.d.) and Singh and Nand (2023), particularly in low-data or real-time scenarios.

c. Histogram Equalization and Contrast Enhancement

Variations in lighting and contrast are common challenges in real-world facial image datasets. Histogram equalization is a technique that redistributes pixel intensity values to improve the global contrast of images, making features more distinguishable.

More advanced methods like Local Contrast Normalization (LCN) and Global Contrast Normalization (GCN) further refine the contrast enhancement process by adjusting intensity variations within localized regions or across the entire image, respectively. These techniques have been shown to improve facial feature visibility, aiding recognition tasks, as discussed by Dalal et al. (2023) and Tribuana et al. (2024).

d. Noise Reduction and Edge Detection

Noise in facial images—arising from compression artifacts, sensor inconsistencies, or environmental factors—can negatively impact model performance. To address this, noise reduction filters such as Gaussian blurring are commonly applied to smooth out irrelevant variations while preserving key structures.

In parallel, edge detection methods like the Canny algorithm are used to highlight the contours of facial components, such as the jawline, eyebrows, and lips. These edges serve as informative features for both traditional and deep learning-based classifiers. The utility of these approaches has been validated in facial analysis studies by Dalal et al. (2023) and Tribuana et al. (2024).

B. Deep Learning-Based Preprocessing Techniques

Deep learning has reshaped the landscape of data preprocessing by integrating tasks like feature extraction directly into model architectures. This approach enables end-to-end training, allowing systems to learn meaningful representations directly from raw inputs—particularly beneficial in emotion recognition where manual feature engineering is often insufficient.

a. Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) have become a backbone for extracting visual features in facial emotion recognition. They automatically learn patterns from facial regions that correlate with emotional states such as arousal and valence. Lightweight models like MobileNet-v2 are especially effective in real-time or embedded applications due to their computational efficiency and reduced model size.

In the MuSe 2023 Challenge, MobileNet-v2 was employed for extracting deep facial features and fine-tuned for valence-arousal prediction tasks. The network demonstrated competitive performance, proving effective at capturing emotional information from facial images. This approach benefited from transfer learning and adaptation strategies, optimizing the model for emotion-specific cues while avoiding overfitting on small datasets (Joudeh et al., 2023; Creţu, 2024).

b. Data Augmentation

Data augmentation is a critical preprocessing strategy in deep learning, particularly for improving generalization in models trained on limited datasets. This technique expands the training data by applying transformations such as random rotations, flips, scaling, brightness adjustment, and the addition of noise. These synthetic variations mimic real-world changes, helping models become robust to diverse input conditions.

In the study by Cruz et al. (2024), data augmentation significantly improved the performance of emotion recognition models, especially in datasets with class imbalance or limited samples. Similarly, Kayaoğlu et al. (2023) demonstrated that applying augmentation techniques led to faster convergence and better emotional state classification accuracy by ensuring broader coverage of emotional variations during training.

c. Distance Weighting Augmentation (DWA)

Distance Weighting Augmentation (DWA) is a novel preprocessing approach that enhances the personalization of emotion recognition models. Unlike traditional augmentation methods, DWA selects similar segments from the dataset based on their proximity in the latent feature space. This targeted

selection helps tailor the training data to a specific user's affective patterns, improving prediction accuracy in personalized models.

Nwadike et al. (2024) introduced this method in the context of the MuSe-Personalisation 2023 dataset, where it was shown to improve the model's ability to detect subtle emotional variations. Their personalized model, trained with DWA, achieved a combined Concordance Correlation Coefficient (CCC) of 0.78, outperforming standard approaches. This highlights DWA's potential for building adaptive emotion recognition systems that are sensitive to individual differences.

C. CNN-Based Approaches

Convolutional Neural Networks (CNNs) have established themselves as foundational architectures in facial emotion recognition systems due to their capacity to learn layered representations directly from raw image inputs. Their ability to capture both spatial locality and hierarchical structure enables effective modeling of facial features associated with emotional expression.

a. Preprocessing for CNNs

Prior to feeding images into CNN architectures, certain preprocessing steps are essential to ensure data uniformity and compatibility with model requirements. Common procedures include resizing all images to a standard dimension (e.g., 48×48 pixels), converting color images to grayscale to reduce computational complexity, and normalizing pixel intensities to stabilize the training process.

Meena and Velmurugan (n.d.) emphasized the importance of image normalization and dimensional standardization in achieving consistent feature learning across samples. Likewise, Singh and Nand (2023) discussed how standard preprocessing operations enhance model convergence and reliability, especially in datasets containing facial images with varying resolutions and lighting conditions. These preprocessing practices contribute to better generalization and more stable model performance in emotion recognition pipelines.

b. Feature Extraction and Fusion

CNNs inherently learn to extract both localized features—such as edges, corners, and facial landmarks—and higher-order global features that relate to overall facial expressions. To further enhance performance, many studies employ feature fusion, where representations from different network layers or models are combined to create a richer and more discriminative feature set.

An effective application of this strategy was demonstrated in work based on the RECOLA dataset, where deep CNN-based features were fused with handcrafted (predesigned) features. This combination yielded improved performance in continuous emotion prediction tasks, achieving a root mean square error (RMSE) of 0.0790 for valence estimation (Creţu, 2024; Joudeh et al., 2023). Such hybrid feature approaches leverage both the model's learned abstractions and domain-specific insights, contributing to more robust predictions in affective computing systems.

D. Advanced Preprocessing Techniques

Advanced preprocessing methods go beyond basic transformations to enhance data quality by leveraging domain expertise and task-specific knowledge. These techniques aim to improve the representativeness and discriminative power of the input data, which is especially valuable in complex affective computing scenarios like human-robot interaction (HRI) and video-based emotion prediction.

a. 3D Morphable Models (3DMM)

3D Morphable Models (3DMM) are employed to reconstruct three-dimensional facial geometry from two-dimensional images, providing a more comprehensive representation of facial structure. This approach helps mitigate pose and illumination variations and allows the generation of synthetic facial sequences that fill sparse regions in the emotional space.

In the study conducted by Cruz et al. (2024), 3DMMs were used to create synthetic expressions targeting underrepresented combinations of arousal and valence. These generated sequences improved the diversity of training data, leading to enhanced model robustness and accuracy, particularly in interactive applications like HRI. By supplementing real-world data with realistic 3D reconstructions, the model could better generalize to subtle or rare emotional states.

b. Multimodal Feature Fusion

Multimodal preprocessing involves combining information from multiple input channels—typically visual, audio, and physiological signals—to capture a more holistic view of emotional states. This strategy acknowledges that emotions are often expressed through a combination of facial cues, vocal tone, and physiological responses such as heart rate or skin conductance.

Yu et al. (2024) demonstrated the effectiveness of multimodal fusion for continuous emotion prediction on the Aff-Wild2 dataset, a widely used benchmark for valence and arousal estimation. Their model integrated visual features from facial expressions with auditory and physiological data, resulting in improved accuracy compared to unimodal systems. This approach is particularly beneficial in real-world applications, where relying on a single modality may be insufficient due to noise or missing data.

c. Spatiotemporal Relationship Learning

Capturing the dynamic nature of facial expressions requires an understanding of both spatial patterns and their evolution over time. Spatiotemporal relationship learning techniques aim to model these changes by considering temporal dependencies alongside spatial features.

Approaches such as Temporal Convolutional Networks (TCNs) and Transformer-based architectures have been successfully applied to learn such relationships in video data. According to Yu et al. (2024), these models enable more accurate prediction of emotional dimensions by modeling how facial configurations evolve over time, rather than relying on static snapshots. This results in more context-aware and temporally consistent emotion recognition, which is essential for applications like video-based affect monitoring and behavioral analysis.

E. Hybrid Approaches

Hybrid preprocessing strategies aim to combine the complementary strengths of traditional handcrafted features and deep learning-derived representations. These approaches often utilize ensemble learning or fusion frameworks to integrate diverse feature types and enhance emotion prediction accuracy.

a. Optimizable Ensemble Regression

One effective hybrid method involves fusing deep visual features extracted from CNN models with traditional, handcrafted features that may encode geometric, appearance-based, or statistical facial information. These combined features are then passed through an optimizable ensemble regression model that predicts continuous emotion values, such as arousal and valence.

In a study by Joudeh et al. (2023), this technique was applied to the RECOLA dataset, a benchmark for continuous emotion recognition. By leveraging both learned and engineered features, the method achieved a root mean square error (RMSE) of 0.1140 for arousal and 0.0790 for valence. This hybrid approach benefits from the expressive power of deep learning while retaining the domain-specific insights captured by predesigned features (Creţu, 2024).

b. Ensemble Valence-Arousal Estimation Framework (EVAEF)

The Ensemble Valence-Arousal Estimation Framework (EVAEF) represents another form of hybrid architecture that integrates multiple feature extractors with temporal modeling components. It combines both static visual features (e.g., from CNNs) and dynamic temporal encoders (e.g., LSTMs or Transformers) to learn how emotional expressions evolve over time in videos.

Liu et al. (2023) demonstrated this framework's effectiveness in the 5th Affective Behavior Analysis in-the-Wild (ABAW) competition. EVAEF achieved a mean Concordance Correlation Coefficient (CCC) of 0.6414 on the official test set, outperforming several baseline models. The ensemble design allows the system to aggregate insights from different architectural perspectives, leading to more robust and context-aware predictions in challenging real-world scenarios.

F. Multimodal Fusion and Temporal-Aware Features

Facial expression analysis in video data presents unique challenges due to the dynamic and context-sensitive nature of emotions. Two key strategies have emerged as essential: multimodal fusion, which combines complementary data types, and temporal modeling, which captures the progression of expressions over time.

a. Multimodal Data Fusion

Multimodal data fusion integrates different types of sensory input, such as visual and audio signals, to improve emotion recognition performance. While visual features extracted from facial expressions capture static and dynamic facial cues, auditory features provide additional context through speech prosody, tone, and intonation.

In the study titled "Valence and Arousal Estimation based on Multimodal Temporal-Aware Features for Videos in the Wild" (2022), a model was proposed that combined visual and aural features for affective state prediction. This multimodal approach was evaluated on the Aff-Wild2 dataset, a standard benchmark for real-world emotion recognition. The fusion-based model achieved a mean Concordance Correlation Coefficient (CCC) of 0.601, indicating significant improvement over unimodal baselines.

The synergy between audio and visual cues helped the model to disambiguate complex emotional expressions that may not be visually apparent alone.

b. Temporal-Aware Features

Temporal-aware preprocessing focuses on capturing the evolution of emotional expressions across video frames. Facial emotions are not static; they unfold over time through subtle changes in muscles and expressions. Therefore, temporal modeling is vital for systems that aim to track emotions accurately in real-world video sequences.

Recent work by Yu et al. (2024) explored the use of Long Short-Term Memory (LSTM) networks and Transformer-based encoders to model these temporal dependencies. These architectures are designed to learn from sequential data, enabling the model to understand context, transitions, and trends in emotional behavior over time. When applied to valence and arousal prediction tasks, these temporal models yielded competitive results, reinforcing the importance of time-aware representations in video-based emotion analysis.

Table 1: Comparison of Preprocessing Methods

Method	Description	Effectiveness
Face Detection and Alignment	Localizes and aligns facial landmarks to reduce variability.	Improved model performance on datasets with varying illumination and poses
Grayscale Conversion and Normalization	Reduces dimensionality and normalizes pixel values.	Enhanced model convergence and accuracy in CNN-based models
Histogram Equalization	Adjusts contrast to handle illumination variations.	Improved robustness to real-world datasets
Data Augmentation	Artificially increases dataset size to improve generalization.	Enhanced robustness and accuracy in models with limited training data
Distance Weighting Augmentation	Expands dataset by identifying similar samples at the segment level.	Improved performance of personalized models, achieving a CCC of 0.78
3D Morphable Models (3DMM)	Generates synthetic sequences for underrepresented values in the AV space.	Improved accuracy and robustness in HRI settings
Multimodal Feature Fusion	Combines visual, audio, and physiological features to capture complementary information.	Competitive performance on datasets like AffWild2
Spatiotemporal Relationship Learning	Captures spatial and temporal correlations in facial expressions.	Enhanced accuracy of valence and arousal prediction in video data
Optimizable Ensemble Regression	Integrates deep visual and predesigned features for prediction.	Achieved RMSE of 0.1140 for arousal and 0.0790 for valence on RECOLA
Ensemble Valence-Arousal Estimation Framework (EVAEF)	Combines multiple feature extractors and temporal encoders.	Achieved a mean CCC of 0.6414 on the 5th ABAW competition

3. CONCLUSION

Preprocessing significantly influences the accuracy and generalizability of valence-arousal estimation models. While traditional techniques offer foundational support, advanced methods such as 3D modeling, spatiotemporal learning, and multimodal fusion deliver superior performance. A balanced combination tailored to the dataset and model architecture is key to enhancing affective computing systems.

REFERENCES

1. D. Dalal, D. Talreja, D. Vaidya, et al., "Comparing the Effects of Various Preprocessing Techniques on the Performance of CNN for Facial Emotion Recognition," in Proc. Int. Conf. Adv. Technol. Appl. (ICACTA), Oct. 2023. [Online]. Available: https://doi.org/10.1109/icacta58201.2023.10392689
2. A.-M. Creţu, "Predicting the Arousal and Valence Values of Emotional States Using Learned, Predesigned, and Deep Visual Features," Sensors, vol. 24, no. 14, Art. no. 4398, Jul. 2024. [Online]. Available: https://doi.org/10.3390/s24134398
3. B. Omer Joudeh, A.-M. Creţu, and S. Bouchard, "Optimizable Ensemble Regression for Arousal and Valence Predictions from Visual Features," Sensors, vol. 23, no. 22, Art. no. 10090, Nov. 15, 2023, doi: 10.3390/s232210090.

4. C. A. Cruz, V. Sichev, T. Igarashi, et al., “Data Augmentation for 3DMM-based Arousal-Valence Prediction for HRI,” arXiv preprint arXiv:2310.03430, Sep. 30, 2024, doi: 10.48550/arXiv.2310.03430.
5. M. Navada, J. Li, H. Seder, and S. Sabanovic, “Improving Personalisation in Valence and Arousal Prediction using Data Augmentation,” arXiv preprint arXiv:2404.08042, Apr. 13, 2024, doi: 10.48550/arXiv.2404.08042.
6. U. Nwanneamaka, T. Wogu, and E. Ezenkwu, “Optimizing Facial Expression Recognition through Effective Preprocessing Techniques,” Journal of Computer and Communications, vol. 11, no. 12, pp. XX–XX, 2023, doi: 10.4236/jcc.2023.1112008.
7. B. Egyedya, T. Tsitong, and S. Kratz, "CNN-Based Emotion Recognition using Data Augmentation and Preprocessing Methods," Sensors, vol. 23, no. 20, Oct. 11, 2023, doi: 10.3390/s23020574.
8. D. Trisnha, M. Azizi, and L. Andi, "Image Preprocessing Approaches Toward Better Learning Performance With CNN," JRESTI (Rekayasa Sistem dan Teknologi Informasi), Jan. 13, 2024, doi: 10.23007/jresti. v8i1.5147.
A. N. Naveer, S. M. Adel, and S. S. Ismail, "Facial Image Pre-Processing and Emotion Classification: A Deep Learning Approach," in Proc. 2019 International Conference on Advanced Science and Technology (ICAST), Nov. 2019, doi: 10.1109/ICAST.2019.8920858.
9. E. Singh, M. Sharma, and A. Bansal, "Enhancing Facial Emotion Recognition through Deep Learning and Image Preprocessing," in Proc. 2023 International Conference on Intelligent Systems, Sep. 8, 2023, doi: 10.1109/IS.2023.10263138.
10. S. Alzahrani, A. Ahmad, K. Ahmad, and D. Kumar, "Emotion Detection Based on Faces through Smart Image Preprocessing," Smart Science, vol. XX, no. XX, 2023, doi: 10.2139/ssrn.4681602.
A. Saeed, B. M. Rodriguez, "Leveraging Deep Learning for Data Processing to Improve Discriminative Modeling Capabilities," Sensors, Jun. 7, 2024, doi: 10.3390/s24123768.
11. "Are 3D Face Shapes Expressive Enough for Recognizing Continuous Emotions and Action Unit Intensities?" IEEE Transactions on Affective Computing, Jan. 1, 2023, doi: 10.1109/TAFFC.2023. 3280630.
A. Sychov, "HSEmotion Team at the 6th ABAW Competition: Facial Expressions, Valence-Arousal and Emotion Intensity Prediction," arXiv preprint arXiv:2403.11920, Mar. 18, 2024.
12. M. Haque, H. Rahayu, and M. Gunawan, "Continuous Valence-Arousal Space Prediction and Recognition Based on Feature Fusion," Sensors, vol. 24, no. 6, Mar. 22, 2024, doi: 10.3390/s240610015.
13. N. Nasir, A. Ali, and T. Hussain, "Improving Valence-Arousal Estimation with Computational Feature Learning and Spatial Normalization," IEEE Xplore, Feb. 15, 2024, doi: 10.1109/ICICCS.2024. 00785.
14. J. Wu, G. Zeng, T. Wang, et al., "Multimodal Fusion Method with Spatiotemporal Sequences and Relationship Learning for Valence-Arousal Estimation," arXiv preprint arXiv:2403.12425, Mar. 19, 2024.
15. X. Li, J. Lei, W. Yang, et al., "EVAEF: Ensemble Valence-Arousal Estimation Framework in the Wild," arXiv preprint arXiv:2302.00023, Feb. 2023.
A. Sychov, "Frame-level Prediction of Facial Expressions, Valence, Arousal and Action Units for Mobile Devices," arXiv preprint arXiv:2205.04729, May 25, 2022.
G. Lin, "Valence and Arousal Estimation based on Joint Multimodal-Aware Features for Videos in the Wild," in Proc. IEEE Conf. CVPR Workshops, Jun. 2022, doi: 10.1109/CVPRW56347.2022.00286.