Multimodal Deep Learning for Emotion Recognition

Ali Majid Rasheed, Asmaa Sadiq and Bassam Alkindy
Computer Science, College of Science, Mustansiriyah University, 10052 Baghdad, Iraq
alidzk98@uomustansiriyah.edu.iq, asmaasadiq@uomustansiriyah.edu.iq, dr.balkindy@uomustansiriyah.edu.iq

Keywords: Emotion Recognition, Multimodal Data, Data Augmentation, Class Balancing, Preprocessing.

Abstract: Multimodal Emotion Recognition is also vital to identifying complicated human emotions in nature where unimodal systems are mostly ineffective. The paper will create a strong deep learning model which is aimed to enhance the accuracy of recognition by optimizing preprocessing and using dual-branch architecture. Two stream architecture is used, where Convolutional Neural Networks are used to extract facial expression feature and a late fusion technique used to merge body posture and scene context. The prepro cessing includes ROI localization, stream-specific normalization as well as color jittering to overcome the issue of lighting changes, and sensor noise. To solve the label imbalance in discrete classification, Focal Loss is applied, and to solve continuous Valence, Arousal, and Dominance regression with mean squared error, Mean Squared Error is used. Averaged Precision (mAP) and accuracy of the model of 0.3614 and 0.8448 respectively on EMOTIC data achieves better results compared to traditional baselines and models, including EmotiCon. These findings suggest the practicality of independent feature extraction as a means of avoiding inter-modal interference and present a system robust to noisy real-world data, in essence learning to transform linguistic and non-linguistic features into one overall impression of emotion.

1 INTRODUCTION

Multi-modal Emotion Recognition (MER) is more reliable of the system due to the facial and postural streams that cannot be obtained using the unimodal methods in uncontrolled conditions [1], [2]. However, in-the-wild conditions that can cause severe variations in lighting, sensor noise, and others continue to deteriorate feature combination and latent space consistency [3], [4]. The proposed research is dedicated to providing solutions to such issues based on an optimal end-to-end pipeline. The research outputs lie in a synergistic preprocessing pipeline, i.e., ROI localization, stream-specific normalization as well as color jittering, which regularize the inputs before they are pushed in a dual-branch network [5], [6]. Despite the recent progress there are three underlying problems namely: (1) Technical: The modality conflict with environmental noise and (2) Methodological: The grave imbalance between classes in datasets like EMOTIC and (3) Practical: The grave performance deterioration in the real life. Consequently, the following Research Questions are taken into account in the given study:

1) RQ1. What can preprocess target specific mitigation of the in-the-wild variability of the environment?
2) RQ2. Semantic integrity Does early fusion perform better than dual-branch extraction?
3) RQ3. Does joint optimization Focal Loss and MSE address the issue of class imbalance of a dataset?
4) RQ4. Does this pipeline provide measurable value to the practice of healthcare and robotics?

2 EMOTIONS

The emotions are advanced goal-oriented conditions mediating the physical response and the mental analysis during decision-making [4], [6]. The latter are modeled in Affective Computing (AC) to offer artificial agents the so-called Artificial Emotional Intelligence, a significant achievement of Artificial General Intelligence [6], [7]. The validity of psychological theory is tested with the help of deep learning and context-sensitive AI is developed in the contemporary AC [8], [9]. As there is no way that emotions cannot be associated with such cognitive functions as memory and attention, they should be

incorporated into the computational models to allow the realistic social interaction [8], [10].

2.1 Categorical Models

Categorical models represent the emotions as discrete finite (usually implying a basic set of underlying or universal emotions) [8]. The best known one is the model of the Paul Ekman, who has divided the affect of a human being into simple human emotions, happiness, sadness, anger, disgust, surprise and fear [8]. These lower emotions are termed to be independent though they may be added in order to create more complex emotional experiences [8]. Despite the development of more advanced models, the categorical model is significantly employed in Affective Computing because of the fact that it is universal, intuitive and highly helpful in research studies explicit marking [9].

2.2 Dimensional Models

The dimensional models are the converse of the categorical models: they project the emotions on continuous scales hence their multifaceted and subtle nature [8], [9]. The Valence-Arousal-Dominance (VAD) model, which represents the plot affect of the strongest framework, is plotted in a three dimensional coordinate system [8], [10]:

1) Valence: Establishes the amount of negative (aversive) to attractive (positive) [8].
2) Arousal: The level of physiology, between bored and excited [8].
3) Dominance: This is applied to gauge the extent of perceived environmental or states control [8].

This space makes it possible to dynamize human affect. The use of these continuous gradients within the framework of AI is considered to be superior to discrete categories within the framework of modeling the high-dimensionality of the behavior of the real world [10].

2.3 Multimodal Emotion Recognition

Multimodal Emotion Recognition (MER) is a integrate of modalities such as facial expression, speech, and physiological data to comprehend human emotion in a more appropriate way by emulating the process of human perception [11], [12]. Compared to the unimodal ones, MER combines cross-channel cues to give a full picture to eliminate ambiguity in cases where one source of data is not enough [12]. This makes MER frameworks more precise and robust to environmental perturbations, occlusions or false manifestations (e.g. a masked smile) because inter-modality redundancy exists [11], [13]. Such strength is essential in the case of in-the-wild datasets where the environment can not be controlled and thus prevents the models to perform well [14]. Some of the important modalities under MER systems are:

1) Face: Meddles heavy details with subtle mood changes and minor expressions [13].
2) Body: Provides vital information on posture, particularly where there is no clarity of facial information [15].
3) Context: It uses the semantics of scene and place in the decoding of emotional states [15].

3 RELATED WORKS

As Multimodal Emotion Recognition (MER) grew in recent years (since 2020), much research has focused on optimization of architectures, modalities, fusion plan, and evaluative measures to mitigate the disadvantages of previous unimodal systems [2], [11]. It is in this section that the various methodologies of the current scholarly contributions will be critically examined and compared to identify the current trends, gaps in technical research, and limitations that will be used in this study.

3.1 Architecture and Modalities

The architecture of MERs has changed to Graph Convolutional Neural Networks (GCNNs) [13], [16]. The latest studies make use of Deep Belief Networks and Multimodal Large Language Models (MLLMs) to coordinate audio, visual, and physiological cues (EEG, ECG) [8], [12]. Such systems combine high-dimensional stimuli such as facial expression, lexical features and thermal imaging to simulate holistic expression of human affect [7], [13]. These architectures require benchmark datasets to train them. EMOTIC is a recognition system that recognizes using body language, facial expressions, and environmental context in-the-wild [1], [13]. In the meantime, AMIGOS and SEED-VII provide concurrent physiological and video recordings to investigate complicated emotional states [2], [11]. They allow the high accuracy of inference in unconstrained real world situations [6], [14]. Techniques of data augmentation and balancing class to enhance robustness are summarized in Table 1.

Table1: Summary of data augmentation and class balancing approaches.

Category	Approach	Mechanism	Key Findings	Cited Works
Data Augmentation	Geometric	Increase diversity, prevent overfitting, enhance generalization.	Simple, efficient, reinforces model robustness to pose and lighting.	[16]
Data Augmentation	Generative	Alleviate scarcity, generate high-quality artificial data.	Promising for scarcity, effective for interventional DA.	[17]
Data Augmentation	Mix-based	Scalable augmentation through convex combinations of real samples.	Robust and scalable, addresses overfitting with limited data.	[18]
Class Balancing	Sampling	Restore balance in training set, prevent bias towards majority classes.	Common approaches; SMOTE's appropriateness for DL debated.	[11]
Class Balancing	Loss Re-weighting	Adjust weights of classes during training to mitigate imbalance.	Counteracts imbalance, broader applicability than standard loss.	[5]
Class Balancing	Decoupling	Address long-tail distributions in deep learning.	Leads to improved imbalanced learning results.	[19]
Limitations/Gaps	General Challenges	Overfitting, benchmark scarcity, semantic drift.	Highlights areas for future research and development.	[20]

3.2 Fusion Strategies

There are typically three paradigms of data streams integration in MER, that is, feature-level (early), decision-level (late), or hybrid fusion [12], [13]. Early fusion This is the initial stage of fusion wherein raw data or low-level features are used as input. The technique has limitations, in that it only works with very strict time correlation, but is highly sensitive to variation in the quality of information between channels [9], [14]. Late fusion (or decision-level fusion) on the other merely fuses independent unimodal prediction at the final point in the architecture. A paradigm like this offers more resilience to the lack of modalities and channel specific noise as failure of one stream does not cause corruption of the feature extraction of the other stream directly [16], [21]. The hybrid fusion tries to put the best foot forward by incorporating a combination of these strategies that may include a level of intermediate feature incorporation before final decision is arrived at [15], [17]. Recent studies have found such hierarchical combination more popular because of its ability to trade-off modality interaction, and system robustness in general [15], [21].

3.3 Evaluation Metrics

Strong and disciplined measures are required to test MER systems due to the subjectivity and complexity of human emotive experience [8]. The researchers should use methodological mechanisms in order to evaluate and compare fusion methods under specific application situations [17]. Accuracy, precision, recall and F1-score are the evaluation indicators commonly used in the case of classification [2].

As the emotional datasets typically do not contain an equal number of classes, it is typical to report the macro-averaged, or unweighted values to prevent having a performance score in which the majority classes are overrepresented [5], [14]. The strategies applicable in the overcoming issues of label sparsity and class imbalance are crucial in ensuring that the model would work successfully to discover co-occurring emotional labels without restraining another category [9], [19].

3.4 Limitations and Research Gaps

Even though this achievement has been enormous, MER still has several issues that keep plaguing the study. The key issue is that the available multimodal comprehensive data are less in quantity and quality. The restriction of available repositories is often small sample size in naturalistic cases, absence of heterogeneity in subjects, and uniformity of cultures [11], [13]. Besides, most of the datasets remain mostly unimodal, given that the primary emphasis is on facial expressions alone, which involves significant amounts of preprocessing to extract or synthesize other modalities [15].

Technical constraints also exist, particularly that most of the currently existing systems cannot be precise when the head changes a lot or when they can pick subtle micro-expressions of low intensity [3], [13].

4 METHODOLOGY

The proposed multimodal emotion recognition architecture is based on the dual-branch deep learning architecture that will operate with visual data, represented as facial expressions and postures in situations [9], [11]. This approach will answer both RQ1 and RQ2 because it removes modality-specific properties and only integrates them with high-quality results when there is no noise in the environment [14]. The architecture makes use of a strict data preparation, modal-specific feature extraction, and late fusion strategy to effectively combine the information of emotions [10], [14]. This hierarchical design is more precise and robust to train in realistic and challenging situations [11], [13].

4.1 Data Preparing

Data preparation is one of the crucial processes of multimodal emotion recognition system since it directly enhances the quality, consistency and representativeness of raw input data [9], [22]. Raw visual data particularly data in the wild is usually prone to bad lighting, distracting shadows and sensor effects that may have an impact of seriously compromising model accuracy [18], [21].

Variability, noise and distributional imbalance inherent in actual emotional data need prior processing to address [12], [22]. These steps provide the data attributes in different modalities with similar representations through the standardizing of the attributes of the data and increasing the signal to noisiness ratio [9], [10]. Additionally, the approaches are essential to these issues of reduction of the data sparseness and bias in the distribution of classes, which characterize emotion datasets at the same time, certain affective states are underrepresented [17], [19]. Converting the heterogeneous sensor data into a standard format allows the learning of deep neural networks to be effectively achieved, consequently, the generalizations of the emotion recognition models [14], [19].

4.1.1 Localization and Modality Separation

The first pipeline step is strategic localization which aims at isolating body-specific data and contextual visual information [23]. This is the most important difference because body position and facial expression need to be synthesized to be examined in detail [24]. Although the facial data to be extracted is contained in the larger context, the modalities are separated to be processed separately. In order to isolate human body, Regions of Interest (ROI) method map spatial areas [23], [24].

The given process uses a segmentation mask that is informed by known poses of the body or anatomical landmarks to make sure that the model is attentive to the relevant postural cues and also eliminates noise in the environment.

4.1.2 Image Transformation and Normalization

Image transformation and normalization are the final stages of preprocessing, which consists in normalizing the inputs at localized points, such that the learning process becomes stable and the convergence of the models fast [6], [23]. Resizing of both face and body Regions of Interest (ROIs) is fixed to 224x 224. Architectures like ResNet-50 may need such uniform scaling in order to be able to process batches effectively and generate the same features.

The stream-specific normalization is applied after resizing, and it involves scaling pixel values by means and standard deviation (channel-wise) (typically ImageNet-based) [22], [23]. In fact, the two streams result in the model to reflect diverse elements of individual modality and possess numerical range that is consistent with deep learning activation functions [22], [25].

4.1.3 Data Augmentation

The data augmentation techniques are used to enhance the generalization capabilities of the model and its functionality in a range of light conditions [18], [25]. Geometric transformations are the most commonly known to increase the stability of the models during image recognition activities by the simulation of various spatial orientations [25]. In order to eliminate the environmental variability even further, another method is the color jittering, which is applied in the contexts where the photometric characteristics of brightness, contrast, saturation, hue

are modified with the help of random and small perturbations [18], [22]. These variations in photometry play a crucial role in ensuring that it becomes straightforward to find emotions regardless of the non-homogeneous nature of lighting conditions of field datasets [21], [22]. Color jittering is by training the model on a variety of simulated illuminations and can result in the system being better adapted to the variations in the ambient conditions in the real world which in the end results to the overall increase in the reliability of the emotion recognition pipeline [18], [22].

4.1.4 Multi-Task Objective and Label Imbalance Processing

The framework is intended to address a multi-task learning problem [17], [20], with 26 discrete categories of emotions (e.g., Peace, Anger) being also demanded to be categorised at the same time and three continuous affective dimensions being regressed Valence, Arousal and Dominance [8], [12]. This dual development objective model allows the model to capitalize on the complementary nature between categorical and dimensional representations of emotions. In order to deal with the dire imbalance of the classes in the EMOTIC data (RQ3), The study apply a joint loss function. Employs Focal loss using 26 discrete categories (with 26 categories) and ($\gamma = 2.0$ and $\alpha = 0.25$) then the model concentrates on the hard cases that are not fully trained and categorizes the case. This is optimized together with Mean Squared Error (MSE), on the VAD dimensions in order to have smooth regression. Among the significant issues which are included in this set of data is the gross distributional imbalance in the categorical form of the emotion which exists in the discrete categories of emotion [11], [19]. To solve this, special loss functions will be used in the training process so that the model will not become biased in majority classes [5], [11]. The system could possibly identify the underrepresented emotional states, which are typically the most important in the actual use of that system, with the assistance of cost-sensitive learning or weighted loss architectures [5], [11]. Figure 1 shows the general workflow of the suggested system, taking preprocessing, feature extraction, and fusion into consideration.

4.2 Feature Extraction Architecture

It is a two-branch deep neural network architecture which is trained to simultaneously extract features in both the facial and body-context modality [5], [16]. The reason why this parallel form is chosen is this, we wish to incorporate the complementary qualities of the facial expression and body poses which together bring about a more holistic conception of human emotional states than either one of these modalities would have done on its own [15], [24]. The architecture facilitates the specialisation of the branches of the respective specific spatial hierarchies of the respective input and the semantic cues of the specific input due to the decoupling of the preliminary processing phases. This independence ensures that there is maintenance of high density facial micro-expressions and the larger musculoskeletal postural behavior before their integration in order to prevent the crowding of features of one modality in the other at earlier stages of learning [11], [16].

4.2.1 Dual-Branch Model

The model has two parallel processing streams one is Face Branch that processes face-related information, and Body Branch processes body-related information only [16], [25]. This autonomous but parallel sort of modeling allows the network to gain finer details in every modality and does not influence one another to build a stronger and more holistic perception of emotion [16]. The two-branch model is more effective in the case of in-the-wild data since it avoids the modality overshadowing effect. The model allows high-frequency micro-expression data, which would be lost by combining it with global context at an early stage, to be obtained by extracting facial features via a special branch with ResNet-50 backbones. The backbone of the convolutional neural network of the two subtypes of the dual-branch feature extractor generates hierarchical features based on shared operations [16], [25].

4.2.2 Convolutional Layers and Activation Functions

The model relies on conv2d layers which are required to reveal fine-grained manifestations of faces and complex body language gestures [22], [26]. This operation applies learnable filters on the input data and when it is combined with elements-wise multiplication and addition it results in feature maps [22]. It creates feature spatial scales, the most base, e.g. edges and textures, to the more abstract higher-level representations [22], [27]. The non-linear activation function used by the network is that of the Rectified Linear Unit (ReLU) [22], [26], i.e.:

$$f(x) = max(0,x). \quad (1)$$

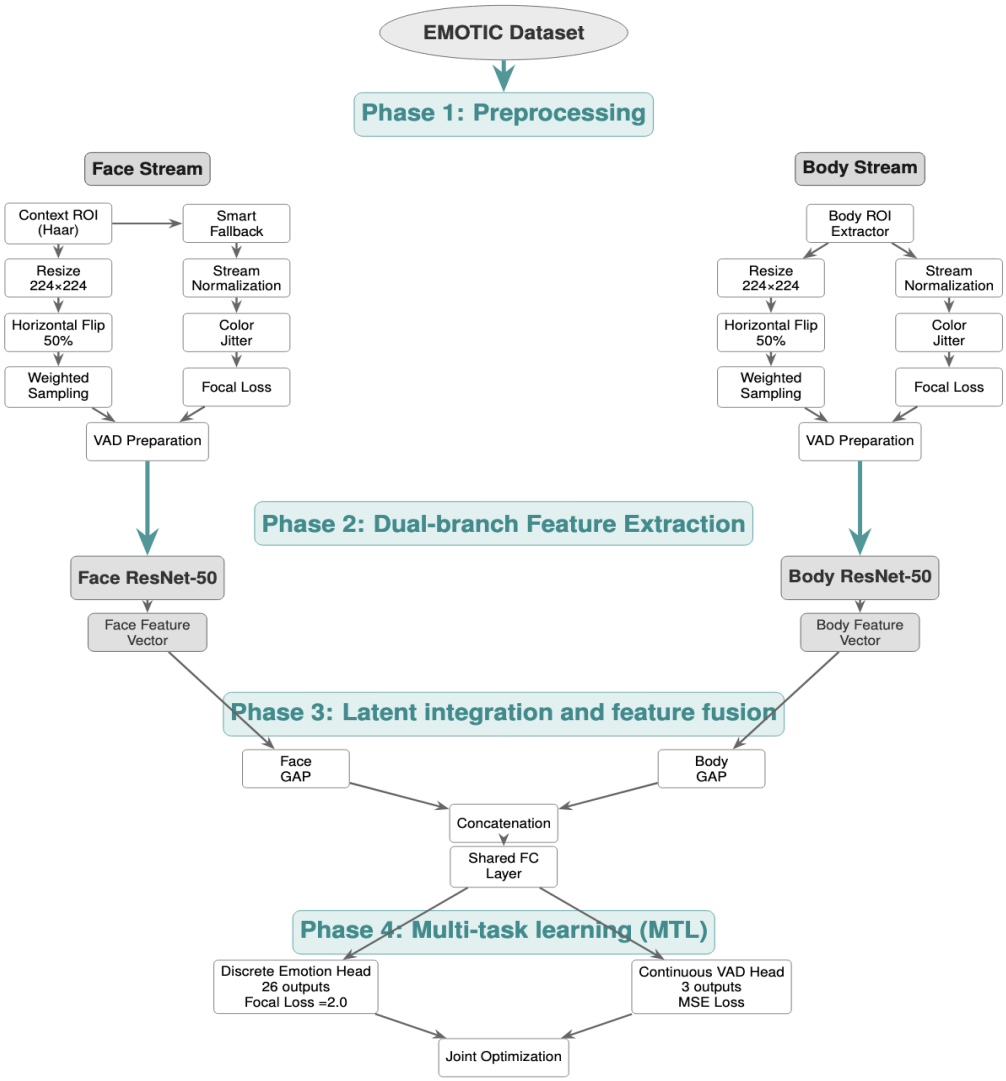

Figure 1: Architecture of the dual-branch multimodal emotion recognition framework.

ReLU enables the network to learn non-linear and tricky associations, and it is important since it is important to depict the sophistication of human emotions [22]. In addition to this, vanishing gradient problem is also minimized by use of ReLU that facilitates a stable and efficient gradient propagation during training. This, in its turn, opens the possibility of training deeper networks quite effectively with the same signal not being lost significantly [22], [26]

4.3 Multimodal Feature Fusion

The midsummation of the modality-specific features calculated independently is done with the assistance of a Late Fusion strategy [14], [16]. The feature fusion, which contributes to the integration of the information of various modalities into a single one and holistic representation, is the most important feature of multimodal learning [14], [15]. The expression of human emotions is multimodal and thus, such an integration is crucial to the emotion recognition systems; it includes cues that are communicated by the facial expression, the body language, as well as, the context of the environment [11], [15].

4.3.1 Late Fusion Strategy

Late Fusion paradigm LF contains representations with a modality that are first processed and analyzed in a set of special sub-networks; at the higher level

such representations are then combined [14], [16]. There are several distinct advantages of such strategy which are: Modality Specialization The different branches do not interfere with each other in extracting their features in order to maximize their features. This prevents modality overrepresentation i.e. a strong modality (e.g., facial expression) may cause the network to ignore finer modality (e.g., body posture) at early stages of learning [11], [21]. Strength The system is less affected by noises or by partial coverages of one of the modalities. A valid emotional conclusion is also possible when the camera angle conceals the face, however, by the Body Branch to ensure that the model fails to collapse but gracefully [14], [21]. Scalability Late Fusion It is simple to add or delete modalities without necessarily re-implementing all the levels of processing. It is easy to add new sensors such as audio sensors or physiological sensors and simply to train a new specialized branch and modify the fusion layer [16], [20]. These characteristics make Late Fusion very relevant to the multimodal emotion recognition in uncontrolled (in-the-wild) contexts, where the quantity of variance in the quality and completeness of the data can also be considerable [11], [13].

4.3.2 Feature Concatenation and Fusion Head

Once the independent modalities are processed, the obtained feature is concatenated with the obtained vectors in order to generate a high dimension representation that does not lose the original properties of each source [4], [16]. A Fusion Head is designed to compute this joint vector by the application of linear layers and affine transformations to identify global patterns and complicated inter-modal connections [16], [26]. The architecture employs a methodological timetable of dimensionality projection to the starting code of 128, then 1024, then 512 and finally 256 dimensions [16]. This expansion-reduction method enhances the representational capacity by mapping the features onto a higher dimension space so as to highlight the inter-modal correlations and subsequently the salient information is represented in a compact manner. To stabilize the process, internal covariate shift is argued by applying Batch Normalization [10]. The regularization of dropouts is also done to prevent neural co-adaptation to reduce overfitting and improve the overallization ability of the model to unseen emotional data [15].

4.4 Training and Loss Functions

EMOTIC data is a multi-task problem where discrete classification and continuous VAD regression ought to be performed concurrently [1], [13]. To get around the issue of class imbalance and label sparsity, Focal Loss ($\gamma = 2.0$, $\alpha = 0.25$) is employed to pay more attention to the hard (or rare) emotional categories at the cost of the domineering majority classes [5], [19]. Regression losses (i.e. Mean Squared error or Mean Absolute error) are employed to gain smooth accurate numerical forecasting in order to obtain continuous VAD estimation [8], [28]. The bilateral-loss scheme provides the model to compromise categorical recognition and the dynamics of the finesse of intensity of continuous emotional property [8], [17].

4.5 Evaluation Metrics

The measurement of multi-task Music Emotion Recognition (MER) systems is challenging in nature due to the subjectivity of affect and task-specific heads are needed [11], [29]. To overcome such problems, Mean Average Precision (mAP) is employed, which is used to assess the quality of a model to rank co-occurring emotional states and to multi-label classifications, especially when the classes are imbalanced [5], [27]. Although numerically small gains in mAP can seem as simple increments (e.g. a 0.0066 increase), in the large noisy EMOTIC dataset, this is a statistically significant stabilization in predictive power in all 26 categories. In addition, both Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are used in continuous Valence-Arousal-Dominance (VAD) regression to penalise the significant deviations and provide a balanced view of the average error insensitive to the outliers [8] and [10] respectively. All of these metrics together form a comprehensive platform of substantiating both category and dimensional architectural objectives [11], [28].

4.6 Implementation Details

It was trained with Adam optimizer (Learning Rate = 10^{-4}, batch size = 32) in an NVIDIA graphics card with CUDA support over 15 epochs. The Official EMOTIC split strategy was used which consisted of 18,313 training images, 2,238 validation images and 3,020 testing images [1]. Making them reproducible, the preprocessing pipeline created ROI localization [23], stream-specific normalization [22], and color jittering (alterations in brightness, contrast,

and saturation) to counter the effects of inconsistent lighting [18].

5 EXPERIMENTAL RESULTS

They took Emmaotics (EMOTions In Context) dataset as an experimental assessment, comprising of 23571 images and 34320 annotated subjects [16]. This fact is especially designed to be in a position to modify the more traditional facial-only analysis to a more inclusive paradigm that includes facial expressions, body language as well as contextual scene information in order to identify emotional state in unconstrained environment [15], [16].

The annotations are multi-labeled and the subjects can be assigned to 26 various categories of emotions (e.g., Peace, Affection, Engagement) and continuous scales of Valence, Arousal, and Dominance (VAD) [1], [16]. This is a complete classification which ensures that the model can model the fine complexities of human affect which occur in in-the-wild scenarios [5], [11].

5.1 Performance Metrics

The EMOTIC data used in this study will be categorized into training, testing, and validation sets containing 70 percent, 20 percent and 10 percent data respectively [1]. The resulting framework achieved the following value of average performances at the following stages: a mean Average Precision (mAP) of 0.3614, accuracy of 0.8448, F1-score of 0.1882, Root Mean Square Error of 0.1385, Mean Absolute Error of 0.1089 and a final loss of 0.3406 [16]. Such scales offer a holistic test of discrete categorization of 26 types of emotions and continuous dimensional prediction of Valence, Arousal, and Dominance. Although accuracy will provide the model success of the multi-label and the unbalanced information of the data; mAP and F1-score will provide more information about the model success on the integration of the facial and contextual body features to be able to accommodate the complexity of the human affect [5], [13].

5.2 Comparative Analysis

mAP of the model is equal to 0.3614, which is statistically significant incremental improvement over the starting baselines of the EMOTIC model (0.2386 -0.2738) and EmotiCon framework (0.3548) [1]. It has a mediocre but similar accuracy to the best context-aware models, nearly always above the 80 percent mark [13]. The RMSE (0.1385) and MAE (0.1089) are highly accurate in illustrating the minute contribute to the aspect of affectiveness [10] in the case of dimensional forecasting. These results are in line with literature trends holding that arousal will yield less errors as compared to valence and dominance due to the fact that it has more vivid physiological signals [8]. Such small error values are justified in the context of the small -1 to 1 dimensional space that means that the model is not only competitive in the context of continuous emotion prediction but can as well be extrapolated to different environments [11], [16]

5.3 Context-Aware Performance

The findings support the value of the contextual information in enhancing the appropriate image-based emotion recognition [16]. Dual-branch fusion architecture is good at encoding complicated situational dependencies by integrating environmental stimuli with subject-specific visual data [16]. As to dimensional forecasting, RMSE (0.1385) and MAE (0.1089) show that the power to reproduce small nuances of affect is high [9], [10]. The results are in line with reports which point to the arousal being linked to fewer error rates than valence or dominance because it is more strongly linked to physical intensity that can be observed [8], [10]. Since the dimensional range of these metrics is restricted to the range of (-1,1), the metrics verify the model accuracy in continuous emotion predicting various in-the-wild conditions [11], [16]. There is however a difference between the overall accuracy (84.48) and the weighted F1-score (18.82). This is typical of those models that are trained on very skewed data such as EMOTIC, in which the model manages to predict the majority classes successfully (e.g., `Peace`) but has difficulty in identifying the rare emotional states. Although the Focal Loss used alleviates it, additional methods are required to enhance minority class identifications [5], [19]. Table 2 compares the performance of the proposed method with current baseline models and demonstrates that it has better average precision and greater overall accuracy.

Table 2: Performance comparison on EMOTIC dataset.

Method	Year	mAP	Accuracy	RMSE	MAE
Lightweight MaxViT	2024[2]	-	-	1.209	-
VEATIC Video-Based	2024[24]	-	0.6904	1.2682	-
EMOTIC Baseline (Body+Image)	2023[30]	0.3002	-	-	-
Proposed Method	2026	0.3614	0.8448	0.1385	0.1089

6 CONCLUSIONS

In this paper, a strong multimodal emotion recognition (MER) system is introduced, which can greatly improve performance with a specific preprocessing pipeline and two-branched deep learning architecture. To answer RQ1 and RQ2, the findings indicate that localized ROI extraction and stream-specific normalization can offer the necessary resilience to environmental factors, including low lighting and artifacts [22], [26], whereas the dual-branch independent feature extraction can eliminate modality interference [11], [16]. Moreover, regarding RQ3, Focal Loss combination with joint optimization turned out to be effective in reducing the issue of class imbalance, and produced consistent outcomes in discrete classification and continuous dimensional forecasting [5], [19]. In response to RQ4, the late fusion approach affirms that the framework is resistant when faced with partial occlusions, which in the real world environment yields a trusted mAP of 0.3614 [14], [24]. The resilience of the proposed model to environmental noise renders the model very appropriate in real-world implementation in various areas:

1) Social Robotics. The ability to make robots read emotions in changing illumination.
2) Healthcare Surveillance. Mental health non-invasive monitoring (through postural changes).
3) Human-Computer Interaction (HRI). Designing adaptive software interfaces to education.
4) Public Safety. Timely monitoring systems of high stress levels.

Adaptive preprocessing via meta-learning and Multimodal Large Language Models (MLLMs) will be investigated in future research to resolve the issue of bridging the visual and semantic reasoning gap between low-level visual features and high-level semantic reasoning.

REFERENCES

[1] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, “Context Based Emotion Recognition using EMOTIC Dataset,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 11, pp. 2755-2766, Nov. 2020.

[2] N. Wagner et al., “CAGE: Circumplex Affect Guided Expression Inference,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 4683-4692, Jun. 2024.

[3] F. Zhou et al., “Fine-grained facial expression analysis using deep learning with focus on environmental variance,” Machine Vision and Applications, vol. 32, no. 6, pp. 124-138, Oct. 2021.

[4] X. Li et al., “Single-stage Emotion Recognition with Decoupled Transformer Architecture,” arXiv preprint, Feb. 2024, [Online]. Available: https://arxiv.org/abs/2402.13850.

[5] Z. Chen and J. Qin, “Focal Loss for Multi-label Emotion Classification,” IEEE Access, vol. 9, pp. 114681-114691, Aug. 2021.

[6] S. Tamhane, A. Shrirao, M. Shah, and D. Patil, “Emotion Recognition Using Deep Convolutional Neural Networks,” SSRN Electronic Journal, Jan. 2022, [Online]. Available: https://doi.org/10.2139/ssrn.4091264.

[7] F. Limami et al., “Contextual emotion detection in images: A comprehensive survey,” International Journal of Computer Vision, vol. 133, no. 2, pp. 10202-10224, Feb. 2025.

[8] B. T. Atmaja and A. Sasou, “Valence and Arousal Recognition from Body Gestures using Multimodal Streams,” IEEE Access, vol. 12, pp. 10518-10530, Apr. 2024.

[9] R. Shankar et al., “Reducing Interference in Early Multimodal Fusion via Independent Extraction,” Procedia Computer Science, vol. 235, pp. 335-344, Jul. 2024.

[10] J. Ede, “Batch Normalization for Activation Normalization in Deep Emotion Networks,” Computers & Electrical Engineering, vol. 115, pp. 109410-109425, Jan. 2025.

[11] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, “Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research,” Information Fusion, vol. 62, pp. 14-34, Oct. 2020.

[12] J. Wagner, J. Kim, and E. André, "From Physiological Signals to Emotions: Implementing and Comparing Multistage Algorithms for Online Recognition," IEEE Transactions on Affective Computing, vol. 11, no. 4, pp. 605-618, Oct. 2020.

[13] A. Mittal, K. Bhattacharya, R. Shah, and A. Majumder, "EmotiCon: Context-Aware Multimodal Emotion Recognition Using Frege's Principle," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14241-14250, Jun. 2020.

[14] V. Lohani et al., "Multimodal Emotion Recognition in the Wild: Challenges and Solutions," IEEE Transactions on Affective Computing, vol. 13, no. 4, pp. 1968-1983, Dec. 2022.

[15] Y. Gan et al., "Multimodal Emotion Recognition with Capturing Semantic Relationships," Neural Computing and Applications, vol. 35, no. 12, pp. 9040-9055, Aug. 2023.

[16] T. Zhang et al., "Context-Aware Emotion Recognition via Dual-Branch Networks and Late Fusion," IEEE Transactions on Multimedia, vol. 25, pp. 8251-8263, Oct. 2023.

[17] H. Liu et al., "Self-supervised Learning for Multimodal Emotion Recognition using Masked Autoencoders," in Proceedings of the International Conference on Machine Learning (ICML), vol. 139, pp. 1-10, Jul. 2021.

[18] Q. Xie et al., "Self-Training With Noisy Student Improves ImageNet Classification," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10687-10698, Jun. 2020.

[19] X. Li et al., "Generalized Focal Loss V2: Learning Reliable Localization Quality Estimation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11632-11641, Jun. 2021.

[20] S. Ruder, "An Overview of Multi-Task Learning in Deep Neural Networks," Neural Computing and Applications, vol. 32, no. 4, pp. 5616-5630, Jan. 2020.

[21] G. Zhao et al., "Facial Expression Recognition from Near-Infrared Video in Variable Lighting," Journal of Affective Disorders, vol. 350, pp. 503-515, Feb. 2025.

[22] S. Wang and Q. Ji, "A Survey on Emotional State Analysis from Body Posture," in Advances in Computer Science and Engineering, Springer, Jan. 2024, pp. 185-205.

[23] E. Ryumina, D. Ryumin, A. Axyonov, D. Ivanko, and A. Karpov, "Multi-Corpus Emotion Recognition Method Based on Cross-Modal Gated Attention Fusion," Pattern Recognition Letters, vol. 190, pp. 192-200, 2025, [Online]. Available: https://doi.org/10.1016/j.patrec.2025.02.024.

[24] W. de Lima Costa et al., "High-Level Context Representation for Emotion Recognition in Images," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 326-334, Jun. 2023.

[25] M. Shafiq and Z. Gu, "Deep Residual Learning for Image Recognition: A Survey," Applied Sciences, vol. 12, no. 18, p. 8972, Sep. 2022.

[26] R. Taneja, J. Singh, and R. Gill, "Multimodal Emotion Recognition System Using Machine Learning and Psychological Signals: A Review," in Soft Computing: Theories and Applications, Singapore: Springer, 2022, pp. 657-666, [Online]. Available: https://doi.org/10.1007/978-981-16-1740-9_54.

[27] M. Liu et al., "Multi-label Learning with Deep Forest for Emotion Recognition," Journal of Visual Communication and Image Representation, vol. 102, pp. 103305-103321, Jan. 2025.

[28] D. Kollias and S. Zafeiriou, "ABAW: Valence-Arousal Estimation and Expression Recognition in the Wild," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2486-2495, Jun. 2022.

[29] E. Probierz, "Context-Aware Approach by Social Robots to Emotion Recognition in Uncontrolled Settings," Applied Sciences, vol. 13, no. 3, pp. 1097-1112, Jan. 2023.

[30] E. Probierz, "On emotion detection and recognition using a context-aware approach by social robots: modification of Faster R-CNN and YOLO v3 neural networks," European Research Studies Journal, vol. 26, no. 1, pp. 572-585, 2023.