Transformer-Based Fusion of Body and Context for Emotion Recognition

Merieme Elkorchi[1,2,*], *Boutaina* Hdioud[1,**], *Rachid* Oulad Haj Thami[1,***], and *Safae* Merzouk[2,****]

[1]Advanced digital enterprise modeling and information retrieval (ADMIR) laboratory, Rabat IT Center, Information retrieval and data analytics team (IRDA), ENSIAS, Mohammed V University in Rabat
[2]SMARTiLab Laboratory, Moroccan School of Engineering Sciences (EMSI Rabat/SMARTILAB)

Abstract. This paper investigates using image-based features, combining body posture and context, to automatically detect emotions in natural settings. We focus on seven emotions: happy, sad, disgust, neutral, surprise, anger, and fear , from the NCAER dataset. We applied the Vision Transformer (ViT-B16) model to each cue. When body features are extracted using the Vision Transformer (ViT-B16) , we achieve an overall accuracy of 58.16%. Similarly, for context feature extraction, we hide the face and body of the main actor in the visual scene and then apply the Vision Transformer (ViT-B16) to the remaining context-related elements, achieving an accuracy of 55.19%. However, by fusing both body and context features, we improve the accuracy to 56.67%, surpassing the GLAMOUR_Net method.

1 Introduction

Emotion recognition plays a crucial role in improving human-computer interaction across a variety of domains. From healthcare [1] to driving monitoring [2], education [3], and the broader field of human-computer systems [4], the ability to understand emotional cues holds significant value. When systems can accurately identify emotions, they gain a deeper understanding of user intent, which ultimately makes interactions more intuitive and meaningful.

The majority of research in emotion recognition has concentrated on facial expressions. These expressions provide powerful signals that machines can learn to interpret. Deep learning techniques, including Long Short-Term Memory[5], Generative Adversarial Networks, and Convolutional Neural Networks, have proven particularly effective in extracting meaningful features from facial data [6]. Through these methods, systems can enhance their ability to detect and analyze the subtleties of human emotions.

Recent studies [7–9] that use Vision Transformers for facial expression recognition have demonstrated their power in surpassing CNNs in emotion recognition tasks. the Vision Transformer leverage self-attention mechanisms that allow them to capture long-range dependencies within images, making them highly effective at processing complex relationships. Their

*e-mail: merieme_elkorchi@um5.ac.ma
**e-mail: boutaina.hdioud@um5.ac.ma
***e-mail: rachid.ouladhajthami@ensias.um5.ac.ma
****e-mail: s.merzouk@emsi.ma

success in facial expression recognition highlights their power in handling more detailed and broad relationships in visual data. In addition to facial analysis, recent works such as [10] have extended the use of Vision Transformers to body-based emotion recognition, effectively capturing postural dynamics and motion patterns. These approaches confirm the flexibility of transformer architectures in handling diverse visual cues relevant to emotional states.

In parallel, emotion recognition research has moved toward combining multiple visual cues rather than being limited to facial expressions. Several studies have explored this direction by integrating facial and contextual information using CNN-based models, such as MCF_NET [11] and GLAMORNet [12]. While these approaches have shown improvements, they remain limited by the local receptive fields and inductive biases of convolutional architectures. In contrast, Vision Transformers have demonstrated strong results in facial expression recognition and body emotion recognition. Motivated by these findings, we extend the use of transformer-based architectures to body and context cues. Our goal is to capture richer and more expressive features for emotion recognition by fully exploiting the global reasoning capabilities of transformer-based architectures.

This paper introduces a novel approach to emotion recognition that combines body and contextual information through the use of Vision Transformer (ViT-B16). By considering not only facial expressions but also body language and context information, we offer a more comprehensive method for detecting emotions. We created a new pipeline using Vision Transformers ability to capture nuanced interdependencies between body posture and context, to improve the accuracy and robustness of emotion recognition. The following are the key contributions from this paper:

- We propose a dual-stream architecture that processes body and context inputs separately, using Vision Transformer (ViT-B16) as a unified backbone.
- We demonstrate that the Vision Transformer (ViT-B16) significantly outperforms CNN-based architectures for both body and context cues.
- We introduce a refined context construction method that masks the body of the central actor using YOLOv8 segmentation.
- We apply focal loss to address the class imbalance present in the NCAER-S dataset.
- We implement a late decision fusion strategy by averaging predictions from the body and context models.
- We validate our approach through extensive experiments and show that the combined model consistently achieves higher accuracy than single-stream baselines, confirming the advantage of fusion with transformers in emotion recognition tasks.

2 Methodology

2.1 Overview

In this section, we will describe our proposed approach. As shown in figure 1, our architecture processes each image through two separate paths. The first path segments the main actor's body and feeds it to a Vision Transformer (ViT-B16) network, which extracts features, flattens them, applies dropout, and outputs seven emotion scores. The second path masks the actor's body, keeps the scene context, and sends these images through an identical Vision Transformer (ViT-B16) network. After both paths finish, our dual-stream architecture fuses their softmax scores, usually by averaging, to form a final prediction. The output of this merge will result in an emotion being chosen out of a possible seven emotions.

The following subsections will elaborate on each model in detail.

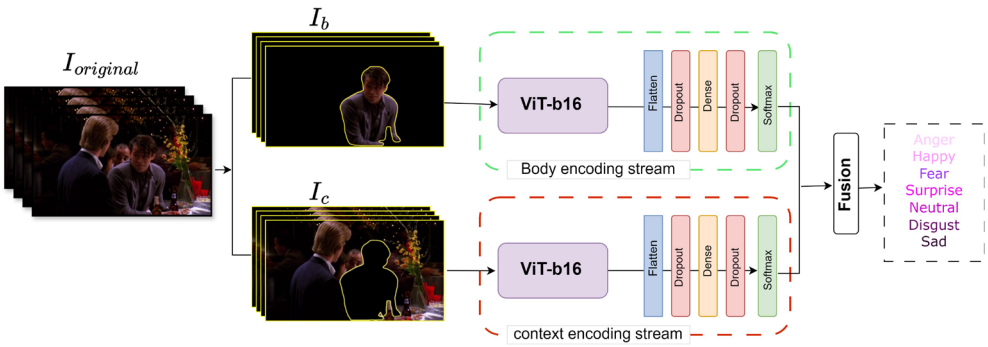

Figure 1. his figure shows a dual-stream architecture based on Vision Transformers (ViT-b16) for emotion recognition.

2.2 Preprocessing pipeline

Before training, we apply a preprocessing step that adds variation across epochs while preserving key features for the model. After that, we resize both body and context images to 112 $\times$ 112 pixels.

2.3 Body encoding stream

We use the human body to capture non-verbal signals. Each image is processed with YOLOv8 [13], which detects all visible bodies. Since we already know the main actor's face, we use its bounding box as a reference. As shown in figure. 2, YOLOv8 returns bounding boxes for every detected body.

To locate the body of the main actor, we compare each body box to the face box. We measure how much they overlap. The body box with the largest overlap is selected as the main actor's body. Once identified, we extract the corresponding mask. We then apply this mask to the original image I_{original} to isolate the body of the main actor. The resulting masked image denoted I_b is the input of the body stream as shown in figure 1.

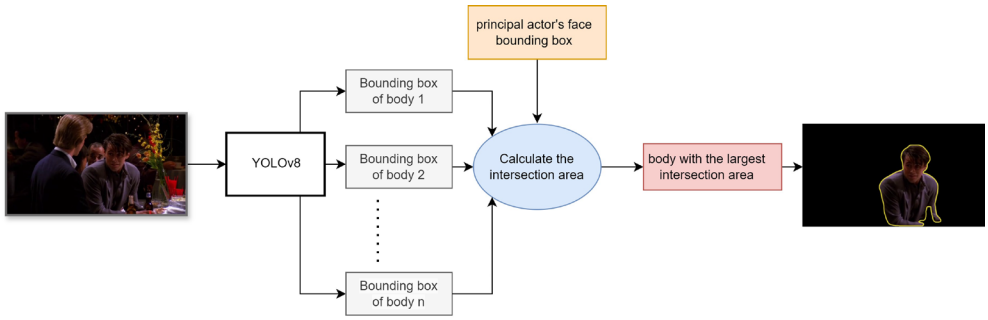

Figure 2. segmentation of the main actor using YOLOv8.

The body encoding module uses a Vision Transformer (ViT-B16) to extract features from the input body image. The model splits the image into small square patches and looks at how

they relate using self-attention. Unlike convolutional networks, Vision Transformer captures both local and global patterns across the entire image.

The output of the transformer is then flattened into one feature vector. Dropout is performed on the output to avoid overfitting, dropping some values during training at random. Then the vector flows through a dense layer with ReLU, helping the model pick up signals tied to emotion.

We add another dropout layer to improve generalization. In the final step, a softmax layer chooses one of seven emotion classes based on the learned features.

2.4 Context encoding stream

To create the context image, we used the actor's body mask, denoted as M_{body}. Since the mask was already available, we set the body area to zero and left the rest of the image unchanged. In simple terms, we applied (1), where M_{body} is the binary mask and I_{original} is the original image. This removed the actor's body while keeping the background, objects, and other people intact. The context stream then focused only on the surroundings, not on the actor.

$$I_{\text{c}} = I_{\text{original}} \times (1 - M_{\text{body}}) \tag{1}$$

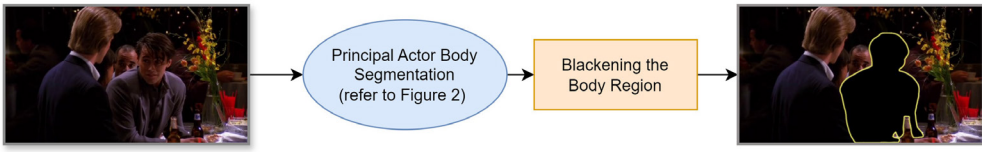

Figure 3. The context image is generated by masking the body region of the main actor

The extracted image, which can be represented as I_c,is used as input to the context encoding stream. We then processed the I_c using the same model architecture as the body module. It passed through a Vision Transformer (ViT-B16), followed by flattening, dropout, a dense layer with ReLU, a second dropout layer, and finally a softmax classifier. This allowed the context stream to extract features from the scene while ignoring the actor.

2.5 Adaptive fusion networks:

We applied adaptive decision fusion to combine predictions from the body and context models. After testing each stream independently, we collected their softmax outputs. We then averaged the outputs from both models to produce a combined prediction. This approach can be expressed as:

$$P_{\text{final}} = \frac{P_{\text{body}} + P_{\text{context}}}{2} \tag{2}$$

where P_{body} and P_{context} are the predicted probability vectors from the two models. To obtain the final class label, we chose the one with maximum probability:

$$\hat{y} = \arg\max(P_{\text{final}}) \tag{3}$$

This late fusion strategy gave equal importance to both sources and allowed the final prediction to reflect information from the body and the context. We then compared the predicted labels with the ground truth to compute the final classification accuracy.

3 Experiments

3.1 Datasets

Our suggested method was tested using the NCAER-S (New CAER-S) dataset. Researchers extracted NCAER-S from CAER-S to make it more robust. CAER-S's training and test sets shared similar video sources, reducing its ability to generalize. NCAER-S resolved this by improving the separation between sets.

figure4 shows the class imbalance in the NCAER-S dataset, with 'Neutral' being dominant and 'Happy' underrepresented. Some test classes contain fewer than 10% of the total samples, which limits the reliability of the evaluation. To improve this, we regrouped all subsets and re-split the data to allocate 10% for testing. While this increased test coverage per class, it did not fully resolve the imbalance, as shown in figure5.

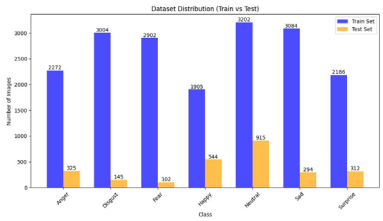

Figure 4. Distribution of images into various classes (Train and Test) in NCAER-S.

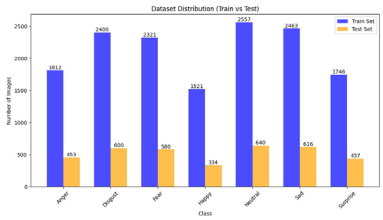

Figure 5. Distribution of images into various classes (Train and Test) in NCAER-S after splitting.

3.2 Loss Function AND Evaluation Metrics

Following [12], we use the standard overall classification accuracy metric to evaluate the results on the NCAER_S dataset and compare them quantitatively with previous approaches.

Our networks are implemented using Tensorflow 2.10 framework [14]. For optimization, we use the Rectified Adam optimization algorithm and Focal Loss function [15], expressed as:

$$\text{FocalLoss} = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\Big[-\alpha\,(1-p_{ic})^{\gamma}\,y_{ic}\,\ln(p_{ic})\Big].$$

Where: N is the number of samples, C is the number of classes, p_{ic} is the predicted probability for sample i and class c, y_{ic} is the correct label, α is a weighting factor to address class imbalance, and γ adjusts how strongly the loss focuses on hard, misclassified examples.

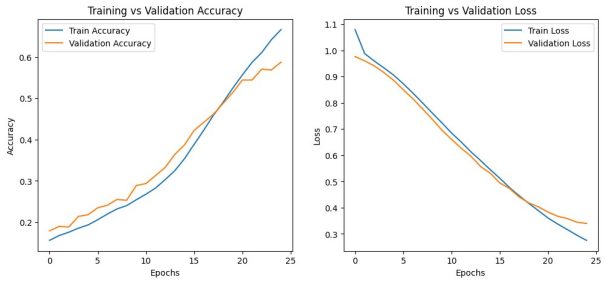

Figure 6. The loss and accuracy curves of our proposed method based on body cues evaluated on the NCAER-S dataset

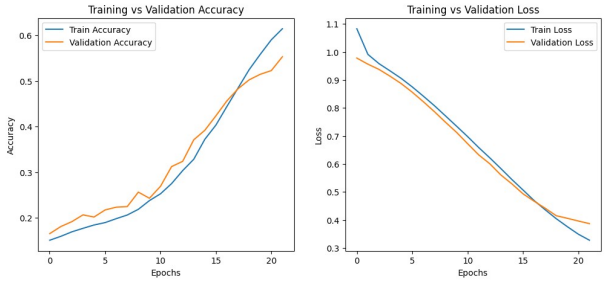

Figure 7. The loss and accuracy curves of our proposed method based on context features evaluated on the NCAER-S dataset

4 Results and Discussion

We analysed the training behavior of two independent models: one on body features and the other on context features, both using the Vision Transformer (ViT-B16) architecture. As shown in figure 6, the loss for the body model decreased consistently over 25 training epochs, starting around 1.0 and reaching approximately 0.32. The context model followed a similar pattern (figure 7), although its final loss remained slightly higher at around 0.38. While both models demonstrated stable convergence, the body model showed a faster and more consistent decrease in loss during training.

These results demonstrate the effectiveness of focal loss in handling class imbalance and guiding the optimization process. It encourages the model to concentrate on challenging or underrepresented instances by reducing the influence of easily classified samples. This mechanism is particularly relevant for our dataset, NCAER-S, which contains unequal distribution of classes. For instance, the 'happy' class includes 1,521 samples, while 'neutral' has 2,557. The sharper decline in the body model's loss suggests that focal loss was especially beneficial in helping the model focus its learning on the most informative examples and achieve stronger generalization.

Table 1. Ablation experiment results of our proposed emotion recognition network.

Method	Body algorithm (YoLov8)	body model	Context model	Acc(%d)
our method	✓	✓	x	58.16%
	✓	x	✓	55.19%
	✓	✓	✓	**56.67%**

4.1 Ablation study

To determine the effect of individual modalities, an ablative study was performed using two separate networks trained only on the body or context inputs. Based on their good performance individually, the predictions were simply averaged, thus creating a fusion network without extra training. In addition, this simple method would ensure a more balanced output by utilizing the benefits of each modality. When we tried fusing the networks using averaging, we found a marked improvement in the classification result, suggesting that the model has successfully learned how to use information from both sets of inputs. This indicates that emotion recognition can be aided by learning distinct cues that appear in body and contextual information. Furthermore, by combining these two modalities, our system is closer to mimicking human emotion perception, as we make use of multiple visual cues in the process of identifying emotions.

4.2 Comparisons to Baseline Methods

In this subsection, we compare our approach against existing baseline methods MCF_NET [11]and GLAMORNet [12]. In Table 2, we see that our dual-stream architecture outperforms all compared models and achieves the highest overall performance. Specifically, it improves classification accuracy by 8.27% over GLAMOR-Net [12].

The results shown in table 2 demonstrate that our dual-stream architecture for emotion recognition, which combines body and context features through two parallel encoding streams, outperforms previous approaches such as MCF_NET [11] and GLAMOR-Net [12]. These models are limited to facial and contextual features and are based entirely on convolutional neural network architectures. In contrast, our method replaces the facial component with body cues and employs the Vision Transformer (ViT-B16) backbone in both streams. In addition, The exclusion of the actor's body from the context image resulted in an observable improvement in classification outcomes. This modification directed the model's attention toward alternative sources of information, such as background configuration, surrounding objects, and other individuals, thereby reducing feature redundancy. Encouraging the model to rely on distinct cues from each input led to a more complementary feature set and enhanced predictive accuracy

our approach introduces two key improvements. First, by leveraging body features instead of facial ones, we capture a broader range of emotional expressions, especially in situations where facial information is ambiguous or occluded. Second, the use of the Vision Transformer (ViT-B16) provides stronger global context modeling than CNNs, allowing the network to better understand spatial relationships and subtle emotional cues distributed across the scene.

The dual-stream architecture, with fine-tuned Vision Transformer (ViT-B16) in both branches, resulted in a clear performance gain. These results confirm that body and context features are complementary and that transformer-based encoding enhances the network's ability to generalize. This suggests that emotion recognition systems benefit from modeling diverse visual cues with architectures capable of long-range feature interaction.

Table 2. Comparison of our method with prior work on the NCAER-S dataset.

Method	Accuracy (%)
MCF_NET [11] (Face + Context)	45.59
GLAMOR-Net (Original) [12] (Face + Context)	46.91
GLAMOR-Net (MobileNetV2) [12] (Face + Context)	47.52
GLAMOR-Net (ResNet-18) [12] (Face + Context)	48.40
Ours (Body + Context)	**56.67**

5 Conclusion

In this work, we proposed a dual-stream framework that combines predictions from both body and context inputs using the Vision Transformer (ViT-B16) architecture for emotion recognition. Our results show that our approach achieves stronger performance than CNN-based models compared to the current state-of-the-art result in the emotion recognition task. Ultimately, the use of focal loss further improved the model's robustness by addressing class imbalance and stabilizing training. Overall, the results on the NCAER-S dataset confirm the effectiveness and reliability of our approach in emotion recognition. In future work, We plan to work on a more challenging and multimodal dataset for emotion recognition. It will include complex backgrounds, multiple faces in a single image, and additional data modalities, allowing us to capture a broader and more diverse set of emotional cues.

References

[1] M. Ali, A.H. Mosa, F. Al Machot, K. Kyamakya, EEG-based emotion recognition approach for e-healthcare applications, in *2016 Eighth International Conference on Ubiquitous and Future Networks (ICUFN)* (2016), pp. 946–950

[2] D. Yang, S. Huang, Z. Xu, Z. Li, S. Wang, M. Li, Y. Wang, Y. Liu, K. Yang, Z. Chen et al., Aide: A vision-driven multi-view, multi-modal, multi-tasking dataset for assistive driving perception, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 20459–20470

[3] H. Farman, A. Sedik, M.M. Nasralla, M.A. Esmail, Facial Emotion Recognition in Smart Education Systems: A Review, in *2023 IEEE International Smart Cities Conference (ISC2)* (2023), pp. 1–9

[4] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, J. Taylor, Emotion recognition in human-computer interaction, IEEE Signal Processing Magazine **18**, 32 (2001). 10.1109/79.911197

[5] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation **9**, 1735 (1997). 10.1162/neco.1997.9.8.1735

[6] S. Singh, F. Nasoz, Facial Expression Recognition with Convolutional Neural Networks, in *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)* (2020), pp. 0324–0328

[7] A. Chaudhari, C. Bhatt, A. Krishna, P.L. Mazzeo, Vitfer: Facial emotion recognition with vision transformers, Applied System Innovation **5** (2022). 10.3390/asi5040080

[8] N.S. Fatima, G. Deepika, A. Anthonisamy, R.J. Chitra, J. Muralidharan, M. Alagarsamy, K. Ramyasree, Enhanced facial emotion recognition using vision transformer models, Journal of Electrical Engineering & Technology pp. 1–10 (2025). https://doi.org/10.1007/s42835-024-02118-w

[9] J. Li, J. Nie, D. Guo, R. Hong, M. Wang, Emotion separation and recognition from a facial expression by generating the poker face with vision transformers, IEEE Transactions on Computational Social Systems **12**, 1548–1562 (2025). 10.1109/tcss.2024.3478839

[10] Y. Xu, J. Zhang, Q. Zhang, D. Tao, Vitpose++: Vision transformer for generic body pose estimation, IEEE Transactions on Pattern Analysis and Machine Intelligence **46**, 1212 (2024). 10.1109/TPAMI.2023.3330016

[11] H. Xu, J. Kong, X. Kong, J. Li, J. Wang, Mcf-net: Fusion network of facial and scene features for expression recognition in the wild, Applied Sciences **12** (2022). 10.3390/app122010251

[12] N. Le, K. Nguyen, A. Nguyen, B. Le, Global-local attention for emotion recognition, Neural Computing and Applications **34**, 21625 (2022). https://doi.org/10.1007/s00521-021-06778-x

[13] G. Jocher, A. Chaurasia, J. Qiu, Ultralytics yolo, https://github.com/ultralytics/ultralytics (2023)

[14] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard et al., {TensorFlow}: a system for {Large-Scale} machine learning, in *12th USENIX symposium on operating systems design and implementation (OSDI 16)* (2016), pp. 265–283

[15] T.Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal Loss for Dense Object Detection, in *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 2999–3007