Multimodal Expression Recognition via Bi-directional Cross-Attention and Temporal Modeling

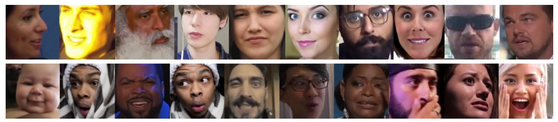
Junhyeong Byeon, Jeongyeol Kim, Sejoon Lim
Kookmin University, Republic of Korea

CVPR JUNE 3-7, 2026 DENVER COLORADO

Abstract

- Expression recognition in unconstrained environments remains challenging due to complex visual and acoustic variations, alongside the dynamic nature of human emotions. To address this, we propose a unified multimodal framework for the EXPR task in the 10th ABAW Challenge.
- Our approach integrates frozen CLIP and Wav2Vec 2.0 encoders, leveraging temporal visual modeling and bi-directional cross-attention to effectively fuse complementary audio-visual cues. Additionally, we apply contrastive learning with label-based text prompts to semantically align multimodal representations with emotion categories.
- Evaluated on the ABAW 10th EXPR benchmark, our framework achieves a Macro F1-score of 0.32, significantly outperforming the official baseline of 0.25. These results validate the importance of temporal modeling and cross-modal interaction for robust, in-the-wild emotion recognition.

10th ABAW Expression Recognition Challenge



Database: Aff-Wild2, a large-scale, in-the-wild audiovisual dataset comprising 548 videos (~2.7M frames).

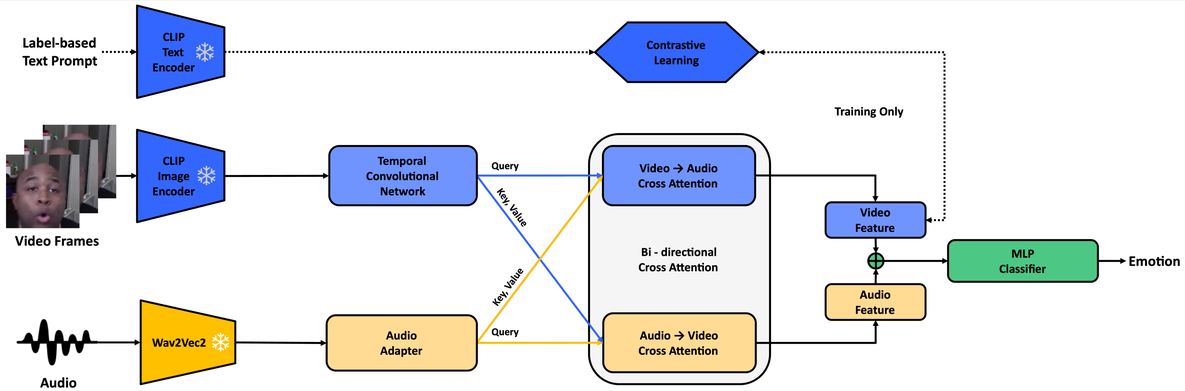
Target Categories: 8 classes (Anger, Disgust, Fear, Happiness, Sadness, Surprise, Neutral, Other).

Challenge Rules: Strictly focused on uni-task expression recognition. Data augmentation and external pre-trained models (not pre-trained on Aff-Wild2) are permitted.

Evaluation Metric: Macro F1-score (average F1 across all 8 categories).

Official Baseline: 0.25 (Pre-trained VGGFACE with MixAugment).

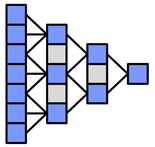
Proposed Method



- Our framework integrates visual and audio modalities for robust emotion recognition.
- We extract visual features using a frozen CLIP1) encoder coupled with a Temporal Convolutional Network2), and process audio through a frozen Wav2Vec 2.03) backbone with an adapter.
- These features are fused via bi-directional cross-attention4) and temporally pooled for final expression classification.
- During training, we apply label-based prompt supervision via contrastive learning to semantically align video representations with emotion categories.

Key Components

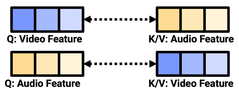
Temporal Modeling



Employs a Temporal Convolutional Network with dilated causal convolutions. By progressively expanding the receptive field, it captures multi-scale dependencies and overcomes the limits of static frames, effectively modeling the dynamic evolution of facial expressions.

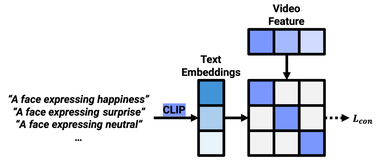
Bi-directional Cross Attention

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



Replaces simple concatenation with deep mutual interaction. By allowing visual and audio features to dynamically query and refine each other, this module deeply integrates complementary cues, creating a robust representation that captures subtle emotional nuances.

Label-based Prompt Supervision



Utilizes CLIP-encoded text prompts (e.g., "A face expressing [Emotion]") as an auxiliary supervision signal. A contrastive learning objective aligns pooled video representations with these text embeddings, guiding the model to learn highly discriminative, category-aware emotion features.

Results

- Our method achieved a 0.36 Macro F1 on the validation set and 0.32 on the test set.

Method	Macro F1
Baseline	0.25
Ours(val)	0.36
Ours(test)	0.32

- We evaluated the class-wise performance on the validation set to analyze the impact of training label distribution.

Label	Train Ratio(%)	F1
Neutral	27.10	0.5521
Anger	3.02	0.1119
Disgust	2.04	0.1846
Fear	1.65	0.1078
Happiness	15.87	0.4917
Sadness	15.43	0.5183
Surprise	5.36	0.3314
Other	29.52	0.6382

Conclusion & Future Work

Conclusion

- Our framework achieved a Macro F1-score of 0.32 on the test set, outperforming the official baseline of 0.25 and ranking 4th on the leaderboard. This demonstrates the effectiveness of the proposed multimodal approach for in-the-wild emotion recognition. The results confirm that integrating complementary audio and visual cues is crucial for handling complex environmental variations.

Future Work

- Future research will focus on addressing class imbalance issues and further refining our multimodal fusion strategies. Through these improvements, we aim to achieve higher performance and develop a more robust system for real-world applications.

Contact

Name: Junhyeong Byeon
Organization: Intelligent Mobility Lab, Kookmin University(Republic of Korea)
Email: Junhyeong0519@kookmin.ac.kr
Website: https://lim.kookmin.ac.kr
Phone: +82-10-5313-6901

References

1) Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PmLR, 2021.
2) Bai, Shaojie, J. Zico Kolter, and Vladlen Koltun. "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling." arXiv preprint arXiv:1803.01271 (2018).
3) Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." Advances in neural information processing systems 33 (2020): 12449-12460.
4) Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).
5) Praveen, R. Gnana, and Jahangir Alam. "Recursive joint cross-modal attention for multimodal fusion in dimensional emotion recognition." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
6) Cabacas-Maso, Josep, et al. "Enhancing Facial Expression Recognition with LSTM Through Dual-Direction Attention Mixed Feature Networks and Clip." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.