Lightweight Models for Emotional Analysis in Video

Quoc-Tien Nguyen, Hong-Hai Nguyen, Van-Thong Huynh[*]
Dept. of ITS, FPT University
Ho Chi Minh City, 71216, Vietnam
{tiennq27,hainh51,thonghv4}@fe.edu.vn

Abstract

In this study, we present an approach for efficient spatiotemporal feature extraction using MobileNetV4 and a multi-scale 3D MLP-Mixer-based temporal aggregation module. MobileNetV4, with its Universal Inverted Bottleneck (UIB) blocks, serves as the backbone for extracting hierarchical feature representations from input image sequences, ensuring both computational efficiency and rich semantic encoding. To capture temporal dependencies, we introduce a three-level MLP-Mixer module, which processes spatial features at multiple resolutions while maintaining structural integrity. Experimental results on the ABAW 8th competition demonstrate the effectiveness of our approach, showing promising performance in affective behavior analysis. By integrating an efficient vision backbone with a structured temporal modeling mechanism, the proposed framework achieves a balance between computational efficiency and predictive accuracy, making it well-suited for real-time applications in mobile and embedded computing environments.

1. Introduction

Human emotion recognition is a subfield of human behavior analysis that has got significant interest from researchers in artificial intelligence, psychology, and human-computer interaction. By utilizing various types of data, such as audio, images, and text, researchers can analyze and predict human emotions as well as continuous actions. Advancements in deep learning has significantly improved the accuracy and efficiency of emotion recognition systems. These technologies can detect emotions in real time, allowing applications across a wide range of fields. Understanding human emotions and behavior can create various applications in areas such as healthcare, autonomous vehicles, robotics, and more [4–6, 16]. For example, emotion recognition is increasingly being used in mental health monitoring, where it helps identify signs of depression or stress based on behavioral patterns. In marketing, consumer emotion recognition enables brands to tailor their products and advertising strategies to evoke desired emotional responses. Moreover, in human-computer interaction, emotion-aware systems are enhancing user experiences by adapting to users' emotional state.

The Emotion recognition has many benefits in life, but emotion recognition still faces various challenges in real-world applications. Human emotions are complex and influenced by various factors such as gender, age, and context, etc. Therefore, the 8th Affective & Behavior Analysis in-the-Wild (ABAW8) workshop [10] specifically addresses the complex challenges inherent in analyzing human affective states and behavioral patterns in unconstrained, real-world scenarios. Unlike controlled laboratory environments, in-the-wild settings present numerous variables including diverse lighting conditions, varying head poses, occlusions, and spontaneous expressions that significantly complicate accurate analysis.

The challenge includes six tasks: Valence-Arousal (VA) Estimation, Expression (EXPR) Classification, Action Unit(AU) Detection, Compound Expression (CE) Recognition, Emotional Mimicry Intensity (EMI) Estimation, Ambivalence/Hesitancy (AH) Recognition. These challenges leverage datasets such as Aff-Wild2, C-EXPR-DB, HUME-Vidmimic2, and BAH, providing a comprehensive benchmark for evaluating affective behavior analysis models.The challenges utilize the Aff-Wild2 [8, 11], C-EXPR-DB [7], HUME-Vidmimic2, and BAH datasets, providing a comprehensive evaluation framework for affective behavior analysis methodologies. In this paper, we solve the tasks such as Valence-Arousal (VA) Estimation, Action Unit(AU) Detection, Emotional Mimicry Intensity (EMI) Estimation, Ambivalence/Hesitancy (AH) Recognition. The VA task predicts valence and arousal in video sequences, while the AU task detects the presence of 12 action units (AUs) in each video frame. The EMI task estimates the intensity of six emotional dimensions (Admiration, Amusement, Determination, Empathic Pain, Excitement, and Joy). Lastly, the

[*] Corresponding author

AH task identifies the presence or absence of ambivalence or hesitancy in each frame.

2. Method

2.1. Visual feature extraction

The visual feature extraction process is facilitated by a small variant of MobileNetV4 [14], pretrained on AffectNet dataset [13]. MobileNetV4 [14] introduces a universally efficient architecture optimized for mobile devices, integrating the Universal Inverted Bottleneck (UIB) and Mobile Multi-Query Attention (Mobile MQA) blocks to enhance feature extraction efficiency. The UIB block unifies key architectural components, including Inverted Bottleneck (IB), ConvNext, Feed Forward Network (FFN), and an Extra Depthwise (ExtraDW) variant, providing flexibility in spatial and channel mixing while improving computational efficiency. Mobile MQA accelerates attention mechanisms by over 39% on mobile accelerators, further optimizing feature representation. Additionally, an advanced neural architecture search (NAS) strategy refines MobileNetV4's design, ensuring mostly Pareto-optimal performance across CPUs, DSPs, GPUs, and dedicated accelerators like Google's EdgeTPU. In this study, we fed a sequence of $224 \times 224 \times 3$ images into the MobileNetV4 to extract multi-scale feature maps from input frames. The backbone is configured to output hierarchical feature representations at different spatial resolutions, allowing for rich semantic extraction. To maintain efficiency while preserving important spatial details, all but the final layers of the backbone are frozen during training. The extracted feature maps serve as input to the subsequent temporal modeling module.

2.2. Temporal aggregation module

To model temporal dependencies in the extracted features, we incorporate a multiscale 3D MLP-Mixer-based [17] to build temporal aggregation module (TAM). This module processes the sequential feature maps using multiple levels of spatial granularity. TAM consist of three mixer layers operate on different feature resolutions: (1) a high-resolution mixer for 28×28 feature maps, (2) a mid-resolution mixer with input size of 14×14, and (3) a low-resolution mixer for highly detail feature maps of size 7×7. Each mixer employs 3D MLP-based transformations to capture temporal relationships while preserving spatial structure. Finally, a fully connected layer maps the aggregated feature representations to the target output space, ensuring effective sequence-level prediction.

3. Experiments and Results

3.1. Dataset

This subsection provides a brief summary of the datasets used in each challenge, such as Action Unit (AU) Detection, Valence-Arousal (VA) Estimation, Emotional Mimicry Intensity (EMI) Estimation.

Table 1. Distribution of AU Annotations in Aff-Wild2

AU	Action	Total Number of Activated AUs
AU 1	inner brow raiser	301,102
AU 2	outer brow raiser	139,936
AU 4	brow lowerer	386,689
AU 6	cheek raiser	619,775
AU 7	lid tightener	964,312
AU 10	upper lip raiser	854,519
AU 12	lip corner puller	602,835
AU 15	lip corner depressor	63,230
AU 23	lip tightener	78,649
AU 24	lip pressor	61,500
AU 25	lips part	1,596,055
AU 26	jaw drop	206,535

Action Unit (AU) Detection This challenge uses a data set that includes 542 videos with annotations for 12 Action Units (AU), which represent facial muscle movements, including brow raisers, cheek raisers, lip tighteners, and jaw drops. The data set consists of 2,627,632 frames captured from 438 unique subjects. Annotations were developed through a semiautomatic methodology that integrates both manual and computational techniques. The data set has been divided into three subsets: a training set (295 videos), a validation set (105 videos), and a testing set (142 videos). Table 1 provides the distribution of the AU annotations of the dataset.

Valence-Arousal (VA) Estimation This challenge utilizes an expanded version of the Aff-Wild2 database, including 594 videos annotated for valence and arousal. The data set consists of 2,993,081 frames captured from 584 subjects. Significantly, sixteen videos contain dual subjects, both of whom received independent annotations. Four expert annotators evaluated the data set following the methodology detailed in [3], continuous valence, and arousal values within the range of [-1, 1].

To maintain subject independence across experimental protocols, the data set has been partitioned into three discrete subsets: a training set (356 videos), a validation set (76 videos), and a testing set (162 videos). This division ensures that individual subjects appear exclusively in one subset.

Emotional Mimicry Intensity (EMI) Estimation. The EMI Challenge considers emotional mimicry through the

Table 2. HUME-Vidmimic2 partition statistics.

Partition	Duration	#Samples
	(HH:MM:SS)	
Train	15:07:03	8072
Validation	9:12:02	4588
Test	9:04:05	4586

HUME-Vidmimic2 dataset, which includes more than 30 hours of audiovisual recordings from 557 participants. This data set was collected in naturalistic environments where participants used their webcams to mimic facial and vocal expressions presented in seed videos, subsequently self-evaluating their mimicry performance on a scale of 0-100. The data set has been systematically partitioned according to the distribution detailed in Table 2, which presents comprehensive statistics for each subset. To facilitate analysis, participants are provided with facial detections extracted from videos using MTCNN [24], processed at a sampling rate of 6 frames per second.

Furthermore, to enable the development of end-to-end methodological approaches [18–22], participants receive pre-extracted features derived from the raw audiovisual data, specifically: facial features processed using Vision Transformer (ViT) [2], audio signals processed using Wav2Vec 2.0 [1].

Ambivalence/Hesitancy Recognition

Action Unit (AU) Detection The $F1$ score uses precision and recall to calculate which ensure a robust assessment of classification performance. The equation of the F_1 score is as follows:

$$F_1 = \frac{2 * precision * recall}{precision + recall} \tag{1}$$

The average F1 score is used to evaluate 12 AUs. The F1 score ranges from 0 to 1, where 1 represents perfect and 0 represents the worst performance. The formula is expressed as follows:

$$F_{\text{AU}} = \frac{\sum_{au} F_1^{au}}{12} \tag{2}$$

Valence-Arousal (VA) Estimation In this task, the Concordance Correlation Coefficient (CCC) [12], $\mathcal{P}$ is used to evaluate the performance of the model. The CCC is a measure that is used to evaluate the agreement between two continuous variables. The CCC range is from [-1,1] where -1 is a negative correlation, 0 is no correlation, and 1 is a high correlation. The formula is expressed as follows:

$$\mathcal{P}_{\text{VA}} = \frac{\mathcal{P}_{\text{V}} + \mathcal{P}_{\text{A}}}{2} \tag{3}$$

where $\mathcal{P}_{\text{V}}$ and $\mathcal{P}_{\text{A}}$ are the CCC of valence and arousal, respectively, which is defined as

$$\mathcal{P} = \frac{2\rho\sigma_{\hat{Y}}\sigma_Y}{\sigma_{\hat{Y}}^2 + \sigma_Y^2 + (\mu_{\hat{Y}} - \mu_Y)^2} \tag{4}$$

where μ_Y was the mean of the label Y, $\mu_{\hat{Y}}$ was the mean of prediction $\hat{Y}$, $\sigma_{\hat{Y}}$ and σ_Y were the corresponding standard deviations, ρ was the Pearson correlation coefficient between $\hat{Y}$and Y.

Emotional Mimicry Intensity (EMI) Estimation. The average Pearson's Correlation Coefficient (ρ)is used to measure the six emotion dimensions:

$$\mathcal{P}_{\text{EMI}} = \frac{\sum_{i=1}^{6} \rho^i}{6} \tag{5}$$

Table 3 shown results of our approach compared to previous studies on Affwild2 dataset [9].

Table 3. Action unit validation set.

Methods	F1 score
Regnet-ViT [23]	0.5280
EmotiEffNet [15]	0.537
Ours with 3 level mixer	0.5369
Ours with 2 level mixer	0.5338
Ours - single mixer	0.5441

References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. 3

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3

[3] Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou, Edelle McMahon, Martin Sawey, and Marc Schröder. Feeltrace: An instrument for recording perceived emotion in real time. In *Proceedings of the ISCA Workshop on Speech and Emotion*, 2000. 2

[4] Albert Haque, Michelle Guo, Adam S Miner, and Li Fei-Fei. Measuring depression symptom severity from spoken language and 3d facial expressions. *arXiv preprint arXiv:1811.08592*, 2018. 1

[5] Varada Kolhatkar, Hanhan Wu, Luca Cavasso, Emilie Francis, Kavan Shukla, and Maite Taboada. The sfu opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus pragmatics*, 4:155–190, 2020.

[6] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2328–2336, 2022. 1

[7] Dimitrios Kollias. Multi-label compound expression recognition: C-expr database & network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5589–5598, 2023. 1

[8] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 1

[9] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021. 3

[10] Dimitrios Kollias, Panagiotis Tzirakis, Alan Cowen, Irene Kotsia, UK Cogitat, Eric Granger, Marco Pedersoli, Simon Bacon, Alice Baird, Chunchang Shao, et al. Advancements in affective and behavior analysis: The 8th abaw workshop and competition. 1

[11] Dimitrios Kollias, Panagiotis Tzirakis, Alan Cowen, Stefanos Zafeiriou, Irene Kotsia, Alice Baird, Chris Gagne, Chunchang Shao, and Guanyu Hu. The 6th affective behavior analysis in-the-wild (abaw) competition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4587–4598, 2024. 1

[12] I Lawrence and Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989. 3

[13] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 2

[14] Danfeng Qin, Chas Leichner, Manolis Delakis, Marco Fornoni, Shixin Luo, Fan Yang, Weijun Wang, Colby Banbury, Chengxi Ye, Berkin Akin, et al. Mobilenetv4: universal models for the mobile ecosystem. In *European Conference on Computer Vision*, pages 78–96. Springer, 2024. 2

[15] Andrey V Savchenko. Hsemotion team at the 6th abaw competition: Facial expressions, valence-arousal and emotion intensity prediction. *arXiv preprint arXiv:2403.11590*, 2024. 3

[16] Gulbadan Sikander and Shahzad Anwar. Driver fatigue detection systems: A review. *IEEE Transactions on Intelligent Transportation Systems*, 20(6):2339–2352, 2018. 1

[17] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021. 2

[18] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of selected topics in signal processing*, 11(8): 1301–1309, 2017. 3

[19] Panagiotis Tzirakis, Stefanos Zafeiriou, and Bjorn W Schuller. End2you–the imperial toolkit for multimodal profiling by end-to-end learning. *arXiv preprint arXiv:1802.01115*, 2018.

[20] Panagiotis Tzirakis, Jiehao Zhang, and Bjorn W Schuller. End-to-end speech emotion recognition using deep neural networks. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5089–5093. IEEE, 2018.

[21] Panagiotis Tzirakis, Jiaxin Chen, Stefanos Zafeiriou, and Björn Schuller. End-to-end multimodal affect recognition in real-world environments. *Information Fusion*, 68:46–53, 2021.

[22] Panagiotis Tzirakis, Anh Nguyen, Stefanos Zafeiriou, and Björn W Schuller. Speech emotion recognition using semantic information. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6279–6283. IEEE, 2021. 3

[23] Ngoc Tu Vu, Van Thong Huynh, Trong Nghia Nguyen, and Soo-Hyung Kim. Ensemble spatial and temporal vision transformer for action units detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5770–5776, 2023. 3

[24] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23 (10):1499–1503, 2016. 3