AUD-TGN: Advancing Action Unit Detection with Temporal Convolution and GPT-2 in Wild Audiovisual Contexts

Jun Yu[1], Zerui Zhang[1], Zhihong Wei [1]*, Gongpeng Zhao[1], Zhongpeng Cai[1], Yongqi Wang[1], Guochen Xie[1], Jichao Zhu[1], Wangyuan Zhu[1], Qingsong Liu[2], Jiaen Liang[2]
[1]University of Science and Technology of China
[2]Unisound AI Technology Co., Ltd.
harryjun@ustc.edu.cn
{igodrr,weizh588,zgp0531,zpcai,wangyongqi,xiegc,
jichaozhu,zhuwangyuan}@mail.ustc.edu.cn
{liuqingsong,liangjiaen}@unisound.com

Abstract

Leveraging the synergy of both audio data and visual data is essential for understanding human emotions and behaviors, especially in in-the-wild setting. Traditional methods for integrating such multimodal information often stumble, leading to less-than-ideal outcomes in the task of facial action unit detection. Addressing these challenges, our study introduces a novel approach that synergistically enhances audio-visual data processing. For audio, we employ Mel Frequency Cepstral Coefficients (MFCC) and Log-Mel spectrogram features, enriched through a pre-trained VGGish network, significantly bolstering the audio feature landscape. Concurrently, in the visual spectrum, we enhance feature extraction using an iResNet model pre-trained on facial datasets, thereby improving the robustness and quality of the visual data representation. With this augmented feature set, Temporal Convolutional Networks (TCN) are applied to meticulously extract and analyze time-series characteristics within each modality, fostering a nuanced understanding of temporal dynamics. The integration of cross-modal information is then achieved through a fine-tuned pre-trained GPT-2 model, facilitating sophisticated and context-aware fusion of the multimodal data. This comprehensive approach not only enhances the accuracy of AU detection but also paves the way for a nuanced comprehension of complex emotional and behavioral expressions in real-world scenarios.

*Corresponding author

1. Introduction

The sixth Competition on Affective Behavior Analysis in-the-wild (ABAW6) [10–21, 42] targets challenges in analyzing human emotions through facial expressions. Facial Action Units (AUs), fundamental for expressing emotions, are the focus of significant research due to their communicative importance [10, 11]. Derived from the Facial Action Coding System (FACS) [29], AUs are critical for a variety of applications, from psychology to security. However, detecting AUs accurately, especially in uncontrolled environments, is complex due to diverse expressions and the necessity for multimodal data integration. Our work responds to this challenge, aiming to refine AU detection methods and explore novel multimodal fusion techniques for a deeper understanding of emotional expressions.

The analysis of facial action units (AUs), essential for interpreting human emotions and expressions, relies on the Facial Action Coding System (FACS) to associate specific AUs with localized facial regions. Traditional methods for detecting AUs utilized handcrafted features to represent these regions[4, 5, 7, 25, 36], laying the groundwork for this field of study. However, these methods were limited in their adaptability to the wide range of facial expressions, often unable to accurately capture the nuances of facial movements or adjust to facial posture changes. The introduction of deep learning marked a new era for AU detection, with deep neural networks offering more sophisticated means for extracting and analyzing facial features. Techniques such as employing face landmarks or segmenting aligned faces into patches have been developed to more accurately locate facial areas related to AUs, yet these approaches often fixedly extracted facial features, limiting their effectiveness. To address these limitations, recent innovations have introduced more flexible and adaptive strategies. For instance, one method employed a three-stage training strategy to enable encoders to adaptively extract features related to facial local regions, although this necessitated additional annotations for face landmarks and depended on multi-task learning[33]. Furthermore, recognizing that the activation of AUs is not isolated but interconnected, recent studies have utilized graph neural networks to explore the relationships between AUs, employing a two-stage training strategy to capture multi-dimensional edge features that reflect the complex web of AU interactions[26]. However, the complexity of such training strategies has highlighted the need for a more streamlined approach to AU detection.

This study initiates with the preprocessing of video data to dissect audio and visual streams, whereupon Log-Mel[27] spectrogram and Mel Frequency Cepstral Coefficients (MFCC)[34] are extracted for audio. Subsequent to this foundational step, we leverage pre-trained VGG[32] and ResNet architectures[8] to distill intricate audio and visual features on a per-frame basis. To circumvent the challenges posed by the temporal continuity of video and the homogeneity among audio frames, our methodology incorporates dilated convolutional layers[23], thereby augmenting the model's capacity for temporal context capture and enriching the extraction of temporal features across distinct input branches. Following feature extraction, concatenation and convolutional operations facilitate the integration of these multimodal inputs. Crucially, the incorporation of a pre-trained GPT-2[22] model, with its sophisticated context-aware attention mechanism, marks a pivotal phase in our approach, enhancing the discernment of nuanced facial expressions and their evolution throughout the video sequence[45]. This rigorously structured framework, which transitions from initial data preprocessing to the application of advanced neural networks, provides a robust strategy for interpreting complex emotional and behavioral cues within video data, underscoring the transformative potential of deep learning in the domain of affective computing. To sum up, our contributions can be summarized as:

- We streamline AU detection by preprocessing video into audio and visual streams, extracting Log-Mel and MFCC features, and utilizing pre-trained VGG and ResNet for advanced feature extraction.
- Our method incorporates dilated convolutional layers to enhance temporal context capture, addressing video's temporal continuity and audio frame homogeneity.
- We employ a pre-trained GPT-2 model for its context-aware attention mechanism, significantly improving the detection and interpretation of nuanced facial expressions throughout video sequences.

2. Related Work

Addressing the complexities of facial action unit (AU) detection, the field confronts notable challenges, including the limited identity variance in prevalent datasets and the extraction of pertinent local features for each AU. Traditional methods have shown substantial limitations [3, 9, 16], particularly those dependent on manual feature specification, due to the intricate and nuanced nature of AU annotations.

To navigate these challenges, recent innovations have incorporated additional facial landmarks to delineate important local features, and have harnessed the power of multi-task learning to refine the efficacy of AU detection models. Notably, the SEV-Net [38] model introduces a mechanism to generate local region attention maps through textual descriptors, offering a fresh perspective to concentrate on salient facial areas crucial for AU analysis. Similarly, Tang et al. [33] advocate a three-stage training strategy that capitalizes on facial landmark information in a multi-task learning framework, thereby directing the model's focus toward pivotal facial regions.

However, these methods typically require supplementary landmark annotations, and may not fully address the intricate web of relationships among AUs. To bridge this gap, Luo et al. [26] have developed a technique leveraging a graph neural network, which employs a two-stage training approach to articulate the relational dynamics between AUs, endeavoring to understand their interconnected behavior. Despite this advancement, their method primarily relies on simple fully connected layers to represent each AU node, which sidesteps the need for additional landmark annotations but necessitates a foundational training period for the network to assimilate node-related information effectively.

Moreover, the quest for a more autonomous learning mechanism that can intuitively grasp critical facial features and their interrelations continues. The exploration extends to devising methodologies that can deduce complex AU patterns and configurations inherently present in facial expressions, with minimal dependency on manual annotations or predefined feature sets. This ongoing research trajectory underscores the field's ambition to craft more intelligent, self-sufficient, and contextually aware AU detection systems, capable of decoding the nuanced spectrum of human facial expressions in a more natural and intuitive manner.

3. Method

In this section, we will describe our proposed approach in detail. As shown in Figure 1, Our methodology for facial action unit (AU) detection commences with preprocessing video into audio and visual streams. For visual features, images $I_v \in \mathbb{R}^{H\times W\times 3}$ are input into a ResNet model[6] pre-trained on Glint360K[1] , producing features $F_v \in \mathbb{R}^{H'\times W'\times C_v}$, where only the final layer is updated.

Audio features are extracted from Log-Mel spectrograms $I_a \in \mathbb{R}^{T\times F}$ through a pre-trained VGGish network, combined with MFCC, resulting in $F_a \in \mathbb{R}^{1\times C_a}$. Temporal dynamics are captured via TCN[2], yielding $T_v \in \mathbb{R}^{1\times C_t}$ for visual and $T_a \in \mathbb{R}^{1\times C_t}$ for audio features. Fusion of these. temporal features through a Transformer network generates a comprehensive representation $F_{fusion} \in \mathbb{R}^{1\times C_f}$, subsequently processed by a multi-class classifier to predict AU presence P_{AU}. This streamlined approach leverages deep learning to efficiently detect AUs, integrating complex audio-visual data.

3.1. Data preprocess

In our data preprocessing pipeline, we meticulously prepare both visual and auditory inputs to ensure that they are optimally primed for feature extraction. For the visual component, each frame of the video sequence is processed through a ResNet network that has been pre-trained on a facial dataset. This pre-training allows the network to generate high-fidelity representations that are particularly attuned to facial features, which are crucial for AU detection. Let I_{frame} denote the input frame and F_{ResNet} represent the output feature vector obtained from ResNet:

$$F_{ResNet} = ResNet(I_{frame}) \in \mathbb{R}^{H'\times W'\times C_v} \tag{1}$$

Here, H' and W' denote the height and width of the processed feature maps, while C_v denotes the number of channels.

Moving on to the auditory aspect, we begin by extracting two types of features: the Mel Frequency Cepstral Coefficients (MFCC) and Log-Mel spectrogram features. These features are particularly effective in capturing the essence of sound and are fundamental to a variety of audio processing tasks. The Log-Mel features are then passed through a VGGish network, which has been pre-trained to encode these features into a robust auditory representation known as VGGish features:

$$F_{VGGish} = VGGish(LogMel(I_{audio})) \in \mathbb{R}^{C_a} \tag{2}$$

In this formula, $LogMel(I_{audio})$ refers to the Log-Mel spectrogram features of the audio input I_{audio}, and F_{VGGish} denotes the encoded VGGish features with C_a being the feature dimensionality. This pre-trained VGGish model serves to effectively distill the audio information into a format that is conducive to our subsequent multimodal analysis, enabling a more comprehensive understanding of the auditory signals associated with the video data.

This dual-faceted preprocessing approach sets a robust foundation for the ensuing stages of our facial action unit detection framework, ensuring that both the visual and auditory modalities are represented with high granularity and are well-suited for the deep learning tasks ahead.

3.2. TCN

In our methodological framework for temporal feature processing, video sequences are segmented into clips each comprising 200 consecutive frames to prepare for Temporal Convolutional Network (TCN) application. The TCN leverages dilated convolutions to process temporal sequences efficiently, enhancing the model's ability to capture broader contextual information without a commensurate increase in computational demand.

Dilated convolutions enable the network to have an exponentially larger receptive field, which is crucial for incorporating long-range temporal dependencies. For an input sequence $X \in \mathbb{R}^{L\times C_{in}}$, where $L = 200$ is the sequence length and C_{in} is the number of input channels, the TCN applies a dilated convolution operation to produce an output sequence $Y \in \mathbb{R}^{L\times C_{out}}$, with C_{out} as the number of output channels. The dilation factor d determines the spacing between the kernel's elements. The dilated convolution operation in TCN, parameterized by weights θ, can be expressed as:

$$Y(t) = (X *_d f)(t) = \sum_{s=0}^{k-1} f(s)\cdot X(t - d\cdot s) \tag{3}$$

where $*_d$ denotes the dilated convolution operation, f represents the filter of size k, and t indexes the time step. The dilation factor d allows the filter to cover a wider span of the input sequence per time step, effectively enlarging the receptive field and enabling the capture of temporal patterns significant for AU detection.

3.3. Leveraging Pretrained Transformer GPT-2

In our study, we employ Temporal Convolutional Networks (TCN) to extract temporal features from each modality, subsequently fusing these into a unified feature vector through concatenation and convolution operations. This process is formalized as follows:

$$F_{concat} = \text{Concat}(T_v, T_a, \ldots) \in \mathbb{R}^{BL\times 3C} \tag{4}$$

where T_v and T_a represent the temporal features derived from visual and auditory modalities, respectively. The concatenated feature vector F_{concat} undergoes a convolutional refinement process to ensure integrated multimodal data processing:

$$F_{conv} = \text{Conv}(F_{concat}) \in \mathbb{R}^{BL\times C'} \tag{5}$$

This convolutional layer optimizes the integrated feature vector F_{concat}, producing F_{conv}, which then serves as input to our GPT-2 based model. · Our framework leverages the pre-trained GPT-2 model, renowned for its advanced feature extraction capabilities. GPT-2's architecture is pivotal, with

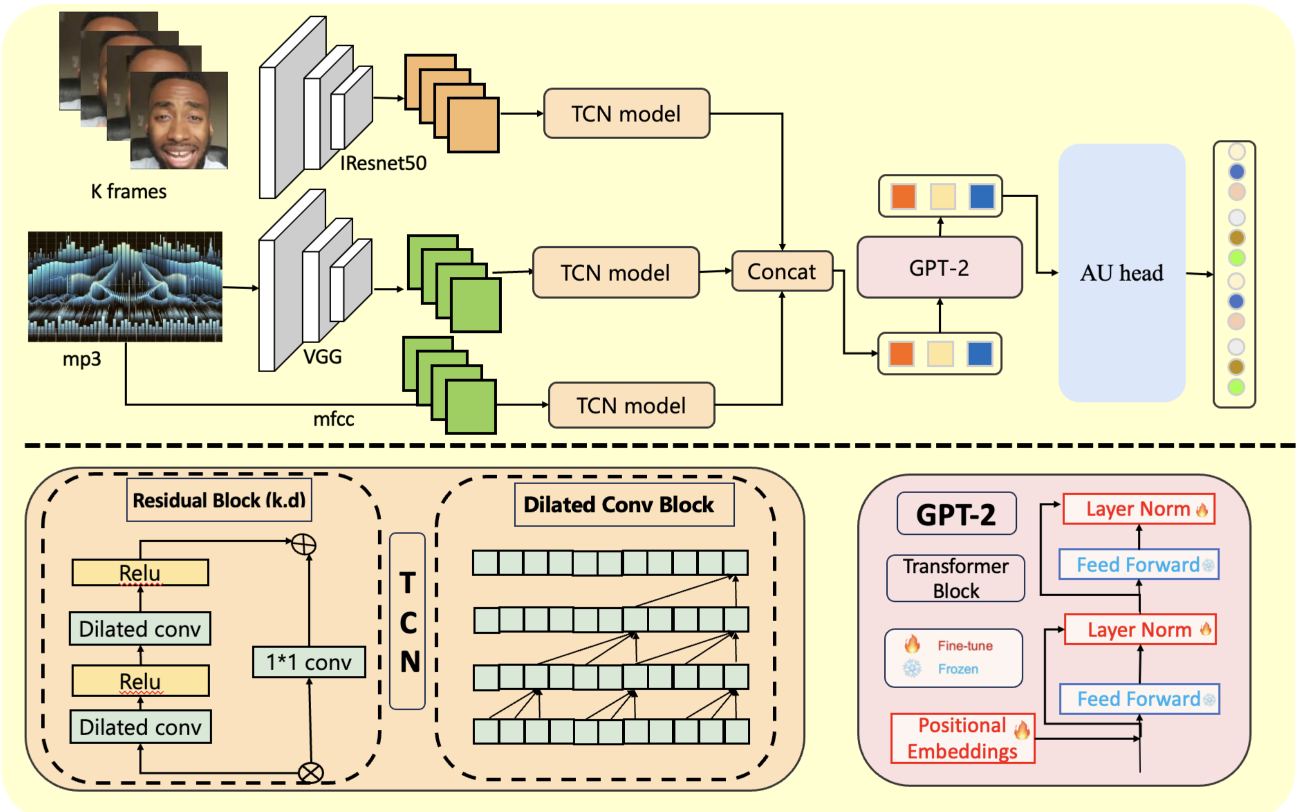

Figure 1. The flowchart presents a multimodal approach for detecting facial action units, employing pre-trained iResnet50 networks for initial feature extraction from video and audio, which are then refined through Temporal Convolutional Networks to capture the temporal dynamics. These features are integrated via a fine-tuned GPT-2 model before being classified by an AU detection head. The detailed submodules illustrate the internal workings of the TCN, emphasizing its dilated convolution blocks for expansive temporal feature capture, and the GPT-2 model, highlighting the transformer mechanism and fine-tuning approach that enables contextual understanding of the features.

multi-headed attention mechanisms, encoding schemes, and position-wise feedforward networks:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_i)W^O \tag{7}$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{8}$$

The feedforward network within GPT-2 is defined as:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{9}$$

Fine-tuning of the GPT-2 model, especially its Layer Normalization (LN) components, is tailored for our specific task of AU detection, utilizing the refined feature vector F_{conv}:

$$F_{GPT2} = GPT2_{\text{LN-finetuned}}(F_{conv}) \in \mathbb{R}^{BL\times N} \tag{10}$$

The resultant feature representation, F_{GPT2}, enriched by the strategic fine-tuning of GPT-2's modules, captures the complex inter-token relationships and temporal dependencies, vital for the nuanced detection of AUs.

3.4. Loss Function

To train our model for the task of facial action unit (AU) detection, we employ the Binary Cross-Entropy (BCE) loss, which for a single instance is defined as:

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\text{BCE}(y_i, \hat{y}_i, W_{au_i}) \tag{11}$$

where N is the number of classes (action units), y_i is the binary label for the i-th AU (1 for presence and 0 for absence), $\hat{y}_i$ is the predicted probability of the i-th AU being present, and W_{au_i} is the weight associated with the i-th AU to address class imbalance.

3.5. Post-Process

Given the structure and composition of the Aff-Wild2 dataset [10, 11, 13–20, 42], it is important to note that the presence of the meaningful label "1" is sparse. This sparsity indicates that such labels are infrequently assigned across the dataset, which poses unique challenges for model training and evaluation. After obtaining the prediction confidence levels for each category, we employ a thresholding technique to segregate the results. The inherent label imbalance in our dataset, predominantly skewed towards the '0' class, results in generally lower confidence scores, not uniformly distributed between 0 and 1. To counteract this and enhance our model's performance, we adjust the threshold for classification. A lower threshold value was empirically found to significantly improve the scoring metrics. Consequently, we systematically explored a range of threshold values on the validation set to identify the optimal solution, thereby optimizing our model's performance in the face of label imbalance.

4. Experiment and Results

In this section, we will provide a detailed description of the used datasets, the experiment setup, and the experimental results.

4.1. Datasets

AU Datasets. The Aff-Wild2 dataset, a substantial extension of the original Aff-wild1 repository, stands at the forefront of affective behavior analysis, offering an unprecedented breadth of annotated data. Spanning 567 videos annotated for valence-arousal dynamics, and 548 videos each for eight distinct expression categories, this dataset encompasses a comprehensive range of human emotions. Additionally, 547 videos meticulously annotated for 12 distinct Action Units (AUs) enhance the dataset's granularity. The dataset is further bolstered by a collection of 172,360 images annotated across the valence-arousal spectrum, six basic expressions plus neutral and 'other' states, and 12 AUs, providing a multifaceted view of human affect.

The Action Unit Detection task, a critical component of the dataset, is represented in 548 videos that capture the six fundamental expressions, the neutral state, and an 'other' category encapsulating affective states beyond the basic emotions. With close to 2.6 million frames and contributions from 431 diverse participants (265 males and 166 females), the dataset's depth is unparalleled, annotated with precision by a team of seven experts. Aff-Wild2 stands as a testament to spontaneous human affect in naturalistic settings, propelling affective computing closer to the complexities of real-world scenarios.

4.2. Training details

In the training phase of our study, we utilized a pre-trained iResNet network, which has been previously validated for its efficacy in previous research[28, 30, 35, 37, 39, 40, 43, 43, 46]. The fine-tuning was confined to the network's final layer parameters, adjusted to a learning rate that is one-tenth of the standard rate. The optimization process was guided by the AdamW optimizer, spanning a duration of 50 epochs.In our training strategy, a crucial element is the optimization of the learning rate schedule. We adopt a linear warmup strategy that begins with an initial rate and linearly increases to reach a learning rate of 0.0001 within the span of 2000 iterations. This gradual increment allows the model to adjust to the complexity of the task, ensuring stable convergence.If there is no improvement on the validation set for five consecutive epochs, the learning rate is scaled down to a tenth of its value. This approach aims to fine-tune the learning process adaptively based on the model's performance. Our model processed video segments with a length of 200 frames each, and we set the batch size to 4. Notably, during the training phase, the 30th epoch marked a milestone as we observed the best performance on the validation set at this point. This methodical training regimen, marked by strategic learning rate adjustments and careful monitoring of validation performance, underscores our commitment to achieving a robust model that reliably understands and classifies affective behaviors as manifested in the Aff-Wild2 dataset.

4.3. Metrics

In evaluating our model's performance for facial action unit (AU) detection, the F1 score is utilized as the primary metric, capturing the balance between precision and recall. Precision (P) and recall (R) are defined as follows:

$$P = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}, \tag{12}$$

$$R = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}. \tag{13}$$

The F1 score is the harmonic mean of precision and recall, providing a single measure that combines the sensitivity and specificity of the model:

$$\text{F1} = 2 \times \frac{P \times R}{P + R}. \tag{14}$$

For a comprehensive assessment, we calculate the mean F1 score (F1_{mean}) across all AU classes:

$$\text{F1}_{\text{mean}} = \frac{1}{N} \sum_{i=1}^{N} \text{F1}_i, \tag{15}$$

where N is the number of AU classes. This metric, F1_{mean}, effectively summarizes the overall performance of

Table 1. Ablation study results on the official validation set, the highest score is indicated in bold.

Method	Pretrained-ResNet	Resnet-Finetune	TCN	GPT-2	Post-Process	F1 Score (%)
baseline						36.5
pretrained	✓					42.8
pretrained+finetune	✓	✓				44.5
TCN	✓	✓	✓			42.6
GPT-2	✓	✓		✓		48.9
TCN+GPT-2	✓	✓	✓	✓		51.4
TCN+GPT-2	✓	✓	✓	✓	✓	**53.7**

Table 2. The average F1 scores (in %) of different teams on the official Aff-wild2 test set. Our results are indicated in bold.

Teams	F1 Score (%)
Netease Fuxi Virtual Human [44]	56.01
CtyunAI [47]	49.41
HSEmotion [31]	48.78
USTC-IAT-United (Ours) [41]	**48.4**
KBS-DGU	46.52
M2-Lab-Purdue [24]	38.32
baseline [21]	36.5

the model, ensuring a robust evaluation of its ability to detect and classify AUs accurately.

4.4. Results

Validation Set and Test Set Results. The average F1 scores (in %) of different teams on the official Aff-wild2 test set are shown in Table. 2. Our method achieves good performance (48.4%) on the official test set, indicating to some extent the good potential of our approach. In addition, we also achieve 53.7% score on the official validation set, more discussion of the validation set results can be found in Sec. 4.5.

4.5. Ablation Study

Pre-trained iResNet: The foundational element in our model's capability to extract complex facial patterns is the pre-trained iResNet. Removing this component led to a marked reduction in performance, with accuracy dropping from 42.8% to 36.5%, underscoring the vital role of iResNet in discerning detailed features crucial for AU detection.

Temporal Feature Extraction with TCN: Temporal dynamics play a pivotal role in facial expression analysis. Implementing TCN for temporal feature extraction significantly improved our model's performance, boosting the accuracy from 44.5% to 46.6%. This increment emphasizes the value of capturing temporal patterns for effective AU detection.

Fine-tuning of the Network's Last Layer: The fine-tuning of the network's last layer to align with our specific dataset nuances increased the performance, raising the accuracy from 42.8% to 44.5%. This enhancement illustrates the fine-tuning's importance in leveraging the network's learned features for the nuanced tasks of AU detection.

Incorporating Pre-trained GPT-2: The integration of a pre-trained GPT-2 model, renowned for its advanced NLP capabilities, resulted in a substantial performance uplift, with accuracy increasing from 44.5% to 48.9%. This improvement showcases the power of sophisticated NLP models in capturing the complex temporal and contextual nuances of facial expressions.

Summary: Our ablation study clearly illustrates the individual and combined impact of key model components on AU detection accuracy. The findings highlight the crucial roles of pre-trained network elements, temporal feature extraction, and model fine-tuning in navigating the intricacies of facial expression analysis in real-world scenarios.

5. Conclusion

In our research, we've demonstrated the effectiveness of integrating Temporal Convolutional Networks (TCN) with pre-trained iResNet and GPT-2 models for the nuanced task of facial action unit (AU) detection in "in-the-wild" settings. By leveraging TCN for dynamic feature extraction and enriching feature representation through pre-trained models, we achieved notable improvements in AU detection accuracy. The results underscore the synergistic impact of combining temporal analysis with advanced neural architectures in enhancing affective computing applications.

6. Acknowledgments

This work was supported by the Natural Science Foundation of China (62276242), National Aviation Science Foundation (2022Z071078001), CAAI-Huawei MindSpore Open Fund (CAAIXSJLJJ-2021-016B, CAAIXSJLJJ-2022-001A), Anhui Province Key Research and Development Program (202104a05020007), Dreams Foundation of Jianghuai Advance Technology Center

(2023-ZM01Z001), USTC-IAT Application Sci. & Tech. Achievement Cultivation Program (JL06521001Y), Sci. & Tech. Innovation Special Zone (20-163-14-LZ-001-004-01).

References

[1] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, et al. Partial fc: Training 10 million identities on a single machine. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1445–1449, 2021. 2

[2] Shaojie Bai, 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

[3] Carlos Fabian Benitez-Quiroz, Yan Wang, Aleix M Martinez, et al. Recognition of action units in the wild with deep nets and a new global-local loss. In *ICCV*, pages 3990–3999, 2017. 2

[4] Wen-Sheng Chu, Fernando De la Torre, and Jeffery F Cohn. Selective transfer machine for personalized facial action unit detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3515–3522, 2013. 1

[5] Xiaoyu Ding, Wen-Sheng Chu, Fernando De la Torre, Jeffery F Cohn, and Qiao Wang. Facial action unit event detection by cascade of tasks. In *Proceedings of the IEEE international conference on computer vision*, pages 2400–2407, 2013. 1

[6] Ionut Cosmin Duta, Li Liu, Fan Zhu, and Ling Shao. Improved residual networks for image and video recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9415–9422. IEEE, 2021. 2

[7] Stefanos Eleftheriadis, Ognjen Rudovic, and Maja Pantic. Multi-conditional latent variable model for joint facial action unit detection. In *Proceedings of the IEEE international conference on computer vision*, pages 3792–3800, 2015. 1

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016. 2

[9] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7680–7689, 2021. 2

[10] 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, 5

[11] Dimitrios Kollias. Abaw: Learning from synthetic data & multi-task learning challenges. In *European Conference on Computer Vision*, pages 157–172. Springer, 2023. 1, 5

[12] 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.

[13] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 5

[14] 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.

[15] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3652–3660, 2021.

[16] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019. 2

[17] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, pages 1–23, 2019.

[18] Dimitrios Kollias, Attila Schulc, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 637–643. IEEE, 2020.

[19] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multi-task learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021.

[20] Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou. Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. *arXiv preprint arXiv:2303.01498*, 2023. 5

[21] Dimitrios Kollias, Panagiotis Tzirakis, Alan Cowen, Stefanos Zafeiriou, Chunchang Shao, and Guanyu Hu. The 6th affective behavior analysis in-the-wild (abaw) competition. *arXiv preprint arXiv:2402.19344*, 2024. 1, 6

[22] Klemens Lagler, Michael Schindelegger, Johannes Böhm, Hana Krásná, and Tobias Nilsson. Gpt2: Empirical slant delay model for radio space geodetic techniques. *Geophysical research letters*, 40(6):1069–1073, 2013. 2

[23] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1091–1100, 2018. 2

[24] Li Lin, Sarah Papabathini, Xin Wang, and Shu Hu. Robust light-weight facial affective behavior recognition with clip, 2024. 6

[25] Mengyi Liu, Shaoxin Li, Shiguang Shan, and Xilin Chen. Au-aware deep networks for facial expression recognition. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–6. IEEE, 2013. 1

[26] Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. *arXiv preprint arXiv:2205.01782*, 2022. 2

[27] Hao Meng, Tianhao Yan, Fei Yuan, and Hongwei Wei. Speech emotion recognition from 3d log-mel spectrograms with deep learning network. *IEEE access*, 7:125868–125881, 2019. 2

[28] Dang-Khanh Nguyen, Ngoc-Huynh Ho, Sudarshan Pant, and Hyung-Jeong Yang. A transformer-based approach to video frame-level prediction in affective behaviour analysis in-the-wild, 2023. 5

[29] Emily B Prince, Katherine B Martin, Daniel S Messinger, and M Allen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 2015. 1

[30] Andrey V Savchenko. Emotieffnet facial features in uni-task emotion recognition in video at abaw-5 competition. *arXiv preprint arXiv:2303.09162*, 2023. 5

[31] Andrey V. Savchenko. Hsemotion team at the 6th abaw competition: Facial expressions, valence-arousal and emotion intensity prediction, 2024. 6

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[33] Yang Tang, Wangding Zeng, Dafei Zhao, and Honggang Zhang. Piap-df: Pixel-interested and anti person-specific facial action unit detection net with discrete feedback learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12899–12908, 2021. 2

[34] Vibha Tiwari. Mfcc and its applications in speaker recognition. *International journal on emerging technologies*, 1(1): 19–22, 2010. 2

[35] Shangfei Wang, Yanan Chang, Yi Wu, Xiangyu Miao, Jiaqiang Wu, Zhouan Zhu, Jiahe Wang, and Yufei Xiao. Facial affective behavior analysis method for 5th abaw competition, 2023. 5

[36] Ziheng Wang, Yongqiang Li, Shangfei Wang, and Qiang Ji. Capturing global semantic relationships for facial action unit recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3304–3311, 2013. 1

[37] Zihan Wang, Siyang Song, Cheng Luo, Yuzhi Zhou, Shiling Wu, Weicheng Xie, and Linlin Shen. Spatio-temporal au relational graph representation learning for facial action units detection, 2023. 5

[38] Huiyuan Yang, Lijun Yin, Yi Zhou, and Jiuxiang Gu. Exploiting semantic embedding and visual feature for facial action unit detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10482–10491, 2021. 2

[39] Yufeng Yin, Minh Tran, Di Chang, Xinrui Wang, and Mohammad Soleymani. Multi-modal facial action unit detection with large pre-trained models for the 5th competition on affective behavior analysis in-the-wild, 2023. 5

[40] Jun Yu, Renda Li, Zhongpeng Cai, Gongpeng Zhao, Guochen Xie, Jichao Zhu, and Wangyuan Zhu. Local region perception and relationship learning combined with feature fusion for facial action unit detection, 2023. 5

[41] Jun Yu, Zerui Zhang, Zhihong Wei, Gongpeng Zhao, Zhongpeng Cai, Yongqi Wang, Guochen Xie, Jichao Zhu, and Wangyuan Zhu. Aud-tgn: Advancing action unit detection with temporal convolution and gpt-2 in wild audiovisual contexts, 2024. 6

[42] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal 'in-the-wild'challenge. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1980–1987. IEEE, 2017. 1, 5

[43] Wei Zhang, Bowen Ma, Feng Qiu, and Yu Ding. Facial affective analysis based on mae and multi-modal information for 5th abaw competition. *arXiv preprint arXiv:2303.10849*, 2023. 5

[44] Wei Zhang, Feng Qiu, Chen Liu, Lincheng Li, Heming Du, Tiancheng Guo, and Xin Yu. Affective behaviour analysis via integrating multi-modal knowledge, 2024. 6

[45] Tian Zhou, Peisong Niu, Liang Sun, Rong Jin, et al. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems*, 36, 2024. 2

[46] Weiwei Zhou, Jiada Lu, Zhaolong Xiong, and Weifeng Wang. Leveraging tcn and transformer for effective visual-audio fusion in continuous emotion recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2023. 5

[47] Weiwei Zhou, Jiada Lu, Chenkun Ling, Weifeng Wang, and Shaowei Liu. Boosting continuous emotion recognition with self-pretraining using masked autoencoders, temporal convolutional networks, and transformers, 2024. 6