MTMSN: Multi-Task and Multi-Modal Sequence Network for Facial Action Unit and Expression Recognition

Yue Jin*
China Pacific Insurance (Group) Co., Ltd.
719 Yishan Road, Shanghai, China
jinyue-007@cpic.com.cn

Tianqing Zheng*
Shanghai Jiaotong University
800 Dongchuan Road, Shanghai, China
sjtuztq@sjtu.edu.cn

Chao Gao
China Pacific Insurance (Group) Co., Ltd.
719 Yishan Road, Shanghai, China
gaochao-027@cpic.com.cn

Guoqiang Xu
China Pacific Insurance (Group) Co., Ltd.
719 Yishan Road, Shanghai, China
xuguoqiang-009@cpic.com.cn

Abstract

Facial action unit (AU) and basic expression recognition are two basic tasks in the area of human affective behavior analysis. Most of the existing methods are developed in restricted scenarios which are not practical for in-the-wild settings. The Affective Behavior Analysis in-the-wild (ABAW) 2021 Contest provides a benchmark for this in-the-wild problem.

In this paper, we propose a multi-task and multi-modal sequence network (MTMSN) to mine the relationships between the above two different tasks and effectively utilize both visual and audio information of the video. We use both AU and expression annotations to train the model and apply a sequence model to further extract associations between video frames. We achieve an AU score of 0.7508 and an expression score of 0.7574 on the validation set.

1. Introduction

Nowadays, analyzing human affect is becoming more and more important for Artificial Intelligence (AI) systems, especially for human-computer interaction. The ability for machines to recognize human faces is mature, but the ability to understand human emotions still has a long way to go.

The Affective Behavior Analysis in-the-wild (ABAW2) 2021 Competition [13][18][20] [16][15][21] [17][30] is held by Kollias et al. in conjunction with ICCV 2021. It provides a benchmark for three main tasks of Valence-Arousal Estimation, seven Basic Expression Classification and twelve Action Unit Detection[8]. The Facial Action Coding System (FACS) is a comprehensive system for describing facial movement. Action Units (AU) are individual components of muscle movements[8].

All three tasks are based on a large-scale in-the-wild database, Aff-Wild2[20][19]. It consists of 548 videos with 2,813,201 frames and provides annotations for all of these tasks. All videos are collected from Youtube which provides a real-world setting. So it is much more difficult to analyze affect than for other datasets. Data imbalance issues make this competition quite challenging. See Fig. 2.

To tackle these problems, we propose a multi-task method to learn Action Unit (AU) and facial expressions jointly using both visual and audio information. First, we train a visual model. Second, we freeze the parameters of the visual model and train the audio model. Third, we concatenate both visual and audio features and train the sequence model.

In summary, the main contributions of this paper are three-fold: (i) We propose a novel framework called Multi-Task and Multi-Modal Sequence Network(MTMSN) for both AU detection and expression recognition, which combines multi-task learning with multi-modal learning to improve the generalization ability; (ii) We use the transformer encoder to extract the sequence features of the video and introduce the attention mechanism; (iii) The proposed framework has the state-of-the-art performance on Aff-Wild2 dataset.

2. Related Work

In the first ABAW competition, lots of teams presented great methods for this challenging problem.

Deng et al.[3] proposed a multi-task learning method to learn from missing labels. They used a data balancing tech-

*These authors contributed equally to this work.

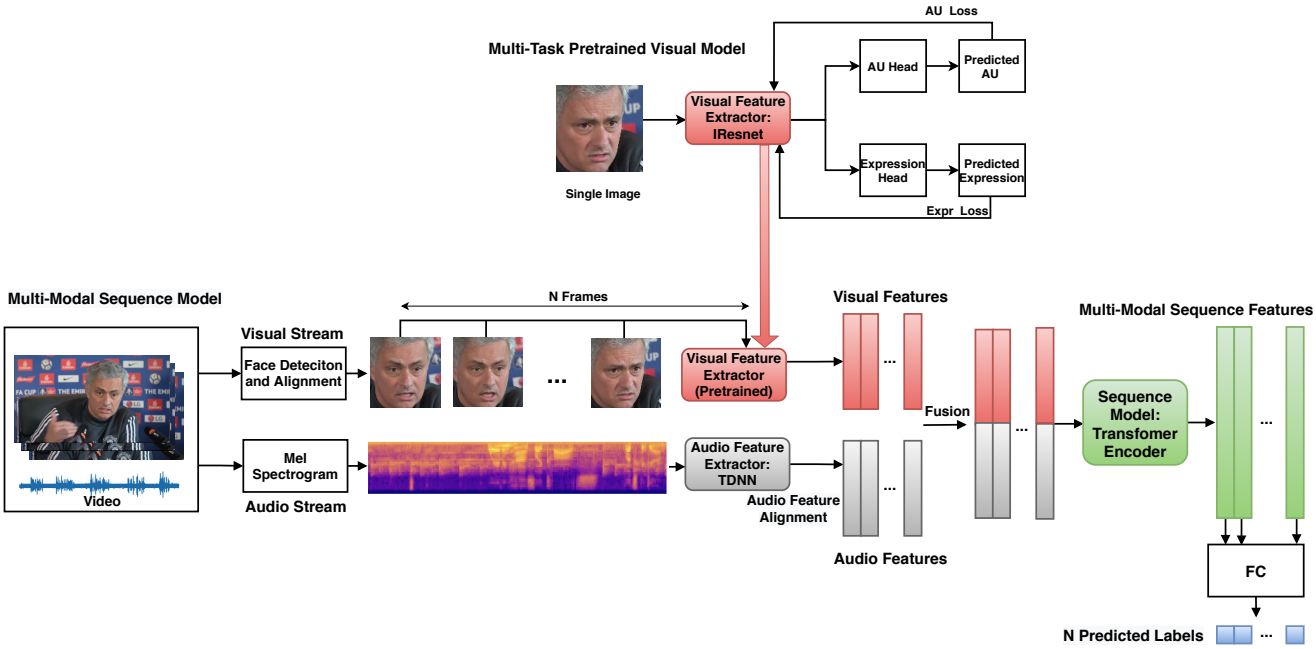

Figure 1. Illustration of the framework of the proposed Multi-Task and Multi-Modal Sequence Network (MTMSN)

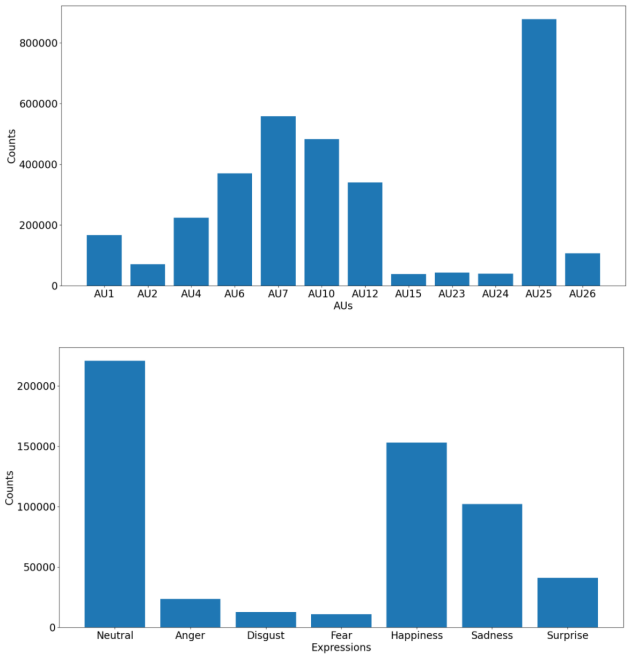

Figure 2. Statistics on the number of AUs and expressions labeled in each category of Aff-Wild2 dataset

nique to the dataset. First, they used the ground truth labels of all three tasks to train a teacher model. Secondly, they used the output of the teacher model as the soft labels. They used the soft labels and the ground truth labels to train their student models.

Kuhnke and Rumberg[22] proposed a two-stream aural-visual model. Audio and image streams are first proposed separately and fed into a CNN network. Then they use temporal convolutions to the image stream. They use additional features extracted during facial alignment and correlations between different emotion representations to boost their performance.

3. Method

We propose a multi-task and multi-modal sequence network (MTMSN) for facial action units and expression recognition. The overall framework can be seen in Fig. 1.

First, we use a multi-task model to train the visual model separately by using both AU and expression annotations. Secondly, we freeze the parameters of the visual stream and add an audio stream to extract the audio features. Finally, the visual features and the audio features are concatenated and fed into a transformer encoder to further extract temporal features.

3.1. Multi-Task Visual model

The Multi-task framework is used to train the visual model as shown in the upper part of Fig. 1.

First, we train the visual backbone with Cosface loss[29]. We choose IResNet100[11] provided by Insightface[9, 2, 4, 6, 10, 7, 5] as the pretrained model. InsightFace efficiently implements a rich variety of popular algorithms of face recognition, face detection, and face alignment and has achieved state-of-the-art performance in

many face recognition benchmarks. Other visual backbones, such as SENet[12] et.al. are also trained with Cosface loss[29]. Using the pre-trained visual backbone boosts the performance because it provides sufficient human face information.

Both AU and expression heads share weights of the same backbone. AU and expression heads are fully connected layers that map features to the number of output classes. The annotations of AU and expression is incomplete, that is, some frames have both AU and expression annotation, but other frames only have one of these two annotations. To tackle this problem, we design two ways to do backpropagation when training.

1) Epoch by epoch: The parameters of the AU head and expression head are updated in rotation epoch by epoch. For example, we have a set of images, annotations of AU, and annotations of expression. At epoch 1, we use images with expression annotations, do backpropagation, and only update the parameters of backbone and expression head. At epoch 2, we use images with AU annotations, do backpropagation, and only update the parameters of the backbone and AU head. Then we repeat the above steps.

2) Batch by batch: The parameters of the AU head and expression head are updated in rotation batch by batch. For example, for the first batch, we use images with expression annotations, do backpropagation, and only update the parameters of the backbone and expression head. For the next batch, we use images with AU annotations, do backpropagation, and only update the parameters of the backbone and AU head. Then we repeat the above steps.

Loss Function

For action unit recognition, we use BCE loss with position weight to tackle the imbalance of positive samples and negative samples. It's possible to trade-off recall and precision by adding weights to positive examples.

$$L_{BCE} = \mathbb{E}[-\sum(w_i t_i \cdot log p_i + (1-t_i)\cdot log(1-p_i))] \quad (1)$$

For expression recognition, we use focal loss[23] to tackle class imbalance problem. It uses a modulating factor to the cross-entropy loss to reduce the loss contribution from easy examples and elevate the importance of hard examples.

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma} log(p_t) \quad (2)$$

3.2. Multi-Modal Sequence Model

The multi-modal sequence model is composed of a visual model, an audio model, and a sequence model. The input of the network is a video and the output is the predicted labels of every frame in the video.

Image frames and the audio stream are extracted from the video and fed into the visual and audio models respectively.

Visual model: The Visual Model is pre-trained using the Multi-task Visual Model in Section 3.1 to extract visual features of a single frame.

Audio model: The Mel spectrogram of the audio is computed using the Torchaudio package and Time-Delay Neural Network (TDNN)[28] is used as the backbone to extract the audio features. TDNN is a widely used network for speech recognition. Peddinti et.al.[25] improved TDNN by using sub-sampling to improve training speed.

Sequence model: The audio features and the visual features are aligned and fused to get the multi-model features. The multi-model features are fed into the encoder layer from transformer[27] to extract the sequence features. The Transformer is a popular sequence network based on attention mechanisms without using RNN layers. It can model the correlation of each feature in the sequence by itself.

Finally, a fully connected layer is used to get the prediction result of each frame.

4. Experiments

4.1. Dataset

The dataset for pretraining visual backbone is Glint360K.[1] It is the largest and cleanest face recognition dataset, which contains 17,091,657 images of 360,232 individuals.

Aff-wild2 dataset[20][19] is used for both AU and expression recognition. We discard annotations with -1.

For AU recognition, we use BP4D[32] as an additional training dataset. This replenishes the number of scarce categories, like AU15, AU23, AU24.

We use the cropped and aligned images provided in the Aff-wild2 dataset.

4.2. Experiment Settings

Our framework is implemented by using Pytorch[24].

Visual model setting: The face recognition model IResNet100[11] provided by Insightface* is used as the pretrained model. Other visual backbones, such as SENet[12] et.al. are also trained with Cosface loss[29] using Glint360K[1] dataset.

We use the cropped and aligned images provided by the Aff-wild2 dataset. The width and height are set to 112 pixels. Data augmentations are random horizontal flip, small random crop, and small random changes to hue, saturation, and lightness. The mini-batch size is set to 64. We use SGD[26] optimizer with momentum and the learning rate is set to 0.001.

*https://github.com/deepinsight/insightface

Table 1. Ablation study of AU recognition results on validation dataset.

Method	Add BP4D[31]	Pos-weight	Multi-task	Seq	Audio	Acc	F1	Score
Baseline[13]	-	-	-	-	-	0.22	0.40	0.31
Visual Model	-	-	-	-	-	0.8856	0.4775	0.6816
Visual Model+	-	✓	-	-	-	0.8621	0.5350	0.6985
Visual Model++	✓	✓	-	-	-	0.8788	0.5328	0.7058
Visual Seq Model	✓	✓	-	✓	-	0.8907	0.5309	0.7108
Multi-modal Seq Model	-	✓	-	✓	✓	0.8786	0.5454	0.7120
Multi-task Visual Model	✓	✓	✓	-	-	0.8960	0.5780	0.7370
Multi-task Visual Seq Model	✓	✓	✓	✓	-	0.9006	0.5815	0.7411
Multi-task Multi-Modal Seq Model	-	✓	✓	✓	✓	0.8987	0.6029	0.7508

Table 2. Ablation study of expression recognition results on validation dataset.

Method	Focal loss	Multi-task	Acc	F1	Score
Baseline[13]	-	-	0.50	0.30	0.366
Visual Model	-	-	0.5793	0.3569	0.4303
Visual Model+	✓	-	0.6288	0.4083	0.4811
Multi-task Visual Model	✓	✓	0.8287	0.7224	0.7574

Audio model setting: We use the following settings to compute a Mel spectrogram of the audio:

- number of mel filter banks $n_{mels} = 64$
- window size $w_{win} = 20ms$
- window stride $t_{stride} = 10ms$

The output dimension of Time Delay Neural Network (TDNN)[25]is set to 512.

Sequence model setting: The input number of frames is set to 30 and the number of encoder layers from the transformer is set to 1. We use SGD[26] optimizer with momentum and the learning rate is set to 0.01.

4.3. Evaluation Metric

For 12 Action Unit Detection, the performance metric[14] is:

$$\mathcal{E}_{total} = 0.5 \times F_1 + 0.5 * \mathcal{T}Acc \tag{3}$$

For 7 Basic Expression Classification, the performance metric[14] is:

$$\mathcal{E}_{total} = 0.67 \times F_1 + 0.33 * \mathcal{T}Acc \tag{4}$$

where F1 Score is the unweighted mean and Accuracy is the total accuracy.

4.4. Results

We use the cropped and aligned images provided in the Aff-wild2 dataset when doing validation. Some video frames are labeled, but there are no corresponding images in the cropped and aligned folder. We discard these frames with their labels when we evaluate our results on the validation set.

Table 3. AU recognition results on the official testing dataset

Method	Average F1	Total Accuracy	Score
Baseline[13]	0.367	0.193	0.28
Ours	0.4892	0.8915	0.6904

Table 4. Expression recognition results on the testing dataset

Method	Average F1	Total Accuracy	Score
Baseline[13]	0.26	0.46	0.326
Ours	0.6834	0.7709	0.7123

Table 1 shows the ablation study of AU recognition results on validation dataset. Table 2 shows the ablation study of expression recognition results on validation dataset. We compare models with different key components. Results show that our multi-task and multi-modal design can effectively lead to better performance.

We achieve an AU score of 0.7508 and an expression score of 0.7574 on the validation set, an AU score of 0.6904, and an expression score of 0.7123 on the testing dataset. Table 3 and Table 4 shows that our method greatly exceeds the baseline.

5. Conclusion and Future Work

We proposed a multi-modal and multi-task learning method by using both visual and audio information for the competition of ABAW2021 in ICCV2021. Our method obtained a score of 0.7508 on AU recognition and 0.7574 on expression recognition using the validation dataset. By using multi-modal information and multi-task training method, the result of our approach far exceeding the baseline result.

For future work, we will investigate the association between AU and expressions in more detail.

References

[1] Xiang An, Xuhan Zhu, Yang Xiao, Lan Wu, Ming Zhang, Yuan Gao, Bin Qin, Debing Zhang, and Ying Fu. Partial fc: Training 10 million identities on a single machine. *arXiv preprint arXiv:2010.05222*, 2020.

[2] Xiang An, Xuhan Zhu, Yang Xiao, Lan Wu, Ming Zhang, Yuan Gao, Bin Qin, Debing Zhang, and Fu Ying. Partial fc: Training 10 million identities on a single machine. In *Arxiv 2010.05222*, 2020.

[3] Didan Deng, Zhaokang Chen, and Bertram E Shi. Multitask emotion recognition with incomplete labels. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 592–599. IEEE, 2020.

[4] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *Proceedings of the IEEE Conference on European Conference on Computer Vision*, 2020.

[5] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

[6] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, 2020.

[7] Jiankang Deng, Anastasios Roussos, Grigorios Chrysos, Evangelos Ververas, Irene Kotsia, Jie Shen, and Stefanos Zafeiriou. The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking. *IJCV*, 2018.

[8] Paul Ekman and Erika L Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[9] Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection. *arXiv preprint arXiv:2105.04714*, 2021.

[10] Jia Guo, Jiankang Deng, Niannan Xue, and Stefanos Zafeiriou. Stacked dense u-nets with dual transformers for robust face alignment. In *BMVC*, 2018.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[13] Dimitrios Kollias, Irene Kotsia, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition, 2021.

[14] Dimitrios Kollias, Irene Kotsia, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. *arXiv preprint arXiv:2106.15318*, 2021.

[15] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 794–800, 2020.

[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.

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

[18] 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.

[19] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*, 2018.

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

[21] 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.

[22] Felix Kuhnke, Lars Rumberg, and Jörn Ostermann. Two-stream aural-visual affect analysis in the wild. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 600–605. IEEE, 2020.

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

[25] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Sixteenth annual conference of the international speech communication association*, 2015.

[26] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[28] Alex Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J Lang. Phoneme recognition using time-delay neural networks. *IEEE transactions on acoustics, speech, and signal processing*, 37(3):328–339, 1989.

[29] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.

[30] 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.

[31] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, and Peng Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–6. IEEE, 2013.

[32] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.