

Multimodal Speech Emotion Recognition via Transformer-Based Hybrid Fusion and Dual Cross-entropy Techniques

Zhixing Yu

Department of Mathematical and Computational Sciences, University of Toronto Mississauga,
Mississauga, Ontario, L5L 1C6, Canada

zhixing.yu@mail.utoronto.ca

Abstract. Speech emotion recognition is gaining increasing interest in the academic sphere due to the advancement of machine intelligence in the service industries. The previous research has already validated the efficacy of multimodality in Speech Emotion Recognition (SER); yet most studies have focused on one-time fusion techniques. This paper proposes a hybrid fusion architecture which optimizes the advantages of multiple fusion techniques and modalities. The model is predominantly based on the rapidly rising Transformer architecture. This study also extends the classic cross-entropy loss and designs a novel loss function which differentiates the misprediction patterns. The architecture is experimented on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset with sufficient cross-validation. It reaches 89.7% accuracy and outmatches the State-of-the-art (SOTA) methods. The performance is further enhanced by the proposed loss function and arrives at 91.1% accuracy. In addition, the models show computation scalability and few needs for hyperparameter fine-tuning. This article concludes that more comprehensive fusion techniques are worth exploration for multimodal speech emotion recognition and Transformers are suitable for emotional characteristics and lead the classification process.

Keywords: Speech emotion recognition, deep learning, Transformer.

1. Introduction

Human emotions are physiological and psychological states which significantly impact human intentions, making emotion recognition a crucial medium for human communication. In the past decade, studies to advance human-computer interaction have witnessed substantial development in speech emotion recognition (SER) [1].

Although the field of SER has been extensively explored, new research interests have emerged fueled by internal and external motivations. Multimodal speech emotion recognition or sentiment analysis showcases better performance over unimodal SER [2, 3]. However, most multimodal SER designs focus on simple structures such as bimodal or trimodal architecture. Moreover, the invention of the Transformer sheds new light on multimodal SER [4, 5]. This sequence-to-sequence architecture excels traditional Long Short-Term Memory (LSTM) for a faster forward path and better Graphics Processing Unit (GPU) compatibility, offering the potential for more complex model engineering and feature processing.

This work constructs a composite multimodal architecture predominantly based on Transformers. It also devises a novel dual cross-entropy (DCE) loss function built upon the classic cross-entropy (CE) loss. The proposed architecture proves to be more accurate than common approaches with single fusion techniques. It is tested on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and outperforms the state-of-the-art methods. In addition, the DCE loss function takes into consideration the distributions of predictions on incorrect classes and further boosts the model's ability to make accurate and stable predictions.

2. Literature Review

This section outlines the research background of the proposed model by presenting related works in multimodal SER, modality fusion, and the relevant applications of the attention mechanism and Transformers.

2.1. Multimodal Speech Emotion Recognition

Multimodal SER refers to conducting emotion recognition through multiple sources of indicators. These include but are not limited to acoustic, visual, text, and electroencephalographical modalities [2, 3]. Traditionally, unimodal emotion recognition prevails, whereas the research focus has been shifted towards multiple modalities for their outstanding effectiveness and data availability [2].

2.2. Modality Fusion

Multimodal SER requires fusion techniques that synthesize the information from different modalities. Such processes commonly involve one of three fusion types: early fusion, late fusion, and intermediate fusion [6]. Early fusion generally applies feature-level fusion, which fuses unprocessed unimodal features. In comparison, late fusion occurs over singular modalities at the decision level after feature processing. Intermediate fusion, a compromised method, allows independent processes of each modality and fuses intermediate results before making unimodal decisions.

Early fusion fully contemplates interdependencies among all features yet potentially overlooks interdependencies of individual modalities. Aziz et al. developed a model using early fusion for emotion analysis [7]. The model exhibits significant performance improvement by a multi-sample dropout method, which relieves dependencies on feature unimodal interrelations [7].

Late fusion maintains modality independence but ignores any interrelation. For instance, Xie et al. tested a two-stage feature selection structure and produced lower accuracies with late fusion than early fusion over four major datasets for emotion recognition [8].

Each fusion technique has its own strengths and weaknesses [9]. However, past model designs for SER primarily adopted only one method and suffered from its limitations.

2.3. Attention- and Transformer-based Emotion Recognition

The attention mechanism is an approach that attracts networks' attention to the relevant parts of the sequence [10]. It has been widely used for modality fusion in emotion recognition [6, 9]. Based entirely on the attention mechanism, the Transformer, a simple but powerful sequence-to-sequence architecture, was developed [4]. Transformers have realized significant achievements in language processing while also promoting remarkable advancements in other fields, including emotion recognition [5]. Both concepts of the attention mechanism and the Transformers are not unfamiliar to the topics of this paper, but only a few apply Transformers for multiple modalities.

3. Method

This section describes the two primary proposals of this article. Section 3.1 illustrates the high-level structure of the proposed hybrid-fusion model. Section 3.2 explains the theory and the properties of the proposed DCE loss function.

3.1. Transformer-based Hybrid Fusion Method

Overall, the architecture involves four modality types: singular modality, intermediate fusion, early fusion, and late fusion. All modalities can carry out their own predictions, while the late fusion modality evaluates the results from the other modalities and produces the final result of the model. These elements are demonstrated in Fig. 1.

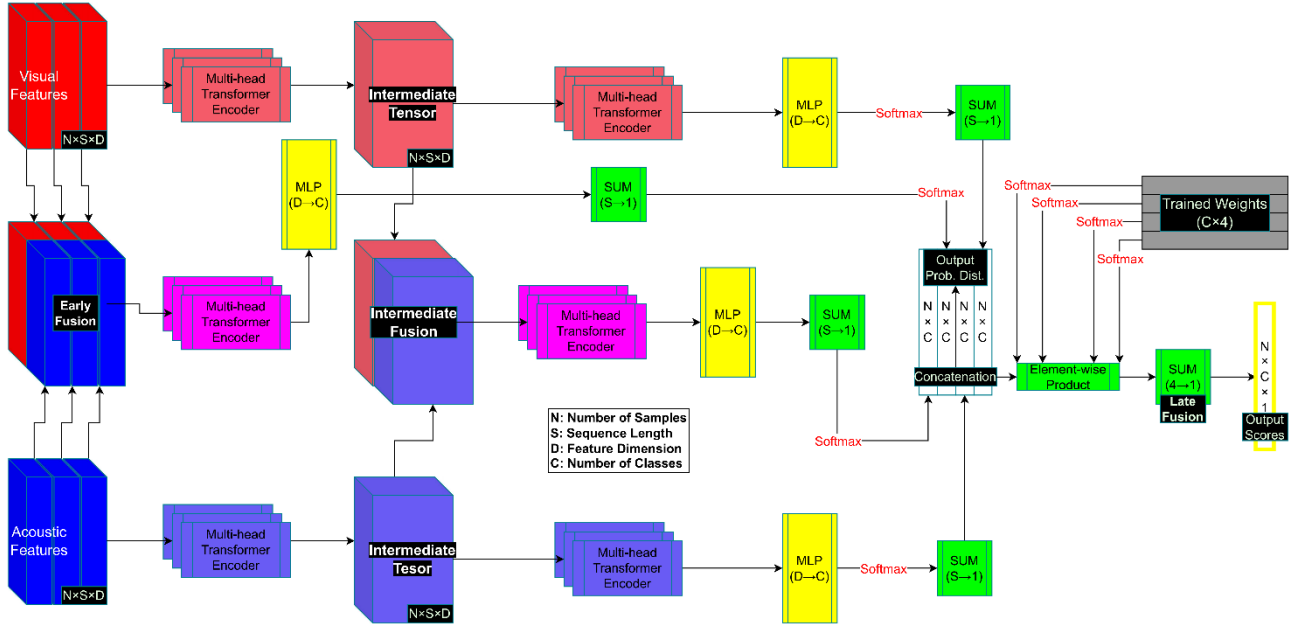


Figure 1. Architecture of the proposed hybrid fusion model (Figure Credits: Original)

The singular modalities are of a homogeneous structure; They are provided with the feature tensors of the shape $N \times S_{mod} \times D_{mod}$ where N is the number of datapoints, S_{mod} and D_{mod} are the sequence length and the number of features specific to individual modalities. The inputs are processed by multi-head Transformer encoder layers; This is followed by a multi-layer perceptron (MLP) layer which reduces the tensor dimension from D_{mod} to the number of classes (C). Finally, the summation over the second dimension yields a tensor of shape $N \times C$, which applies to the softmax function to generate a probability distribution over the emotion classes for each data point.

The early and the intermediate fusion modalities follow the same structures as the singular modalities, except they employ fusion techniques to connect intermodal information. The early fusion occurs on the features of singular modalities while the intermediate fusion occurs on the intermediate tensors between the Transformer encoder layers. Both are open to diverse fusion methods, but the fused tensor is required to maintain a sequential structure to accommodate Transformers.

Late fusion consists of $C \times M$ (number of other modalities) trained weights. The Softmax of these weights along the first axis is applied as element-wise multipliers for the concatenation of the probability distributions outputted by the other modalities. Each weight after Softmax represents the credibility of the corresponding model's prediction on the corresponding label. Each label prediction ranges from 0 to 100%, and the Softmax weights over the corresponding label sum to 1. Thus, the summation over the label axis of their element-wise multiplication ranges from 0 to 1. This results in scores suitable for the multi-margin loss function, which is adopted in the proposed architecture for late fusion [11]. Meanwhile, the performances of the CE loss were also investigated by converting scores into probability distributions with the Softmax function.

3.2. Dual Cross-entropy Loss Function

Apart from the late fusion modality, all modalities of the proposed method output probability distributions from the softmax function. They, therefore, apply to the proposed DCE loss function, which extends the CE loss function. DCE loss adds extra punishment for differences in probabilities between the actual and incorrect labels. Let $\mathbf{x}$ be the predicted probability distribution and y be the index of the actual class, the equation is as follows:

$$\mathcal{L}_{DCE}(\mathbf{x}, y) = \mathcal{L}_{CE} + \sum_{i=0}^{C-1} \max(\mathcal{L}_{CE} + \ln \mathbf{x}[i], 0) \quad (1)$$

Where $\mathcal{L}_{CE} = -\ln \mathbf{x}[y]$.

This function has the following desirable properties compared to the classic CE loss: (1) Punishment for misclassification: It imposes additional punishment if and only if the prediction is

false. This further distinguishes the correct and incorrect predictions, where the former is generally preferred over the latter, even when the predicted probabilities are equal. (2) Higher punishment for lower probability of incorrect prediction: For $\beta > 0$, $\mathbf{x}[y] < \mathbf{x}[i]$, and $\mathbf{x}[i] + \beta \leq 1$, $\ln \mathbf{x}[i] - \ln \mathbf{x}[y] > \ln(\mathbf{x}[i] + \beta) - \ln(\mathbf{x}[y] + \beta)$. This suggests that the DCE loss function disfavors low-probability incorrect predictions, which means discouraging excessive divergence from the correct label. (3) Higher punishment for uniform probability distribution: For $\beta > 0$, $\mathbf{x}[y] \leq \mathbf{x}[i]$ and $\mathbf{x}[i] + 2\beta \leq 1$, $2 \ln(\mathbf{x}[i] + \beta) - 2 \ln \mathbf{x}[y] > \ln(\mathbf{x}[i] + 2\beta) - \ln \mathbf{x}[y]$. Hence, for equal differences in probability, the DCE loss function encourages less uniform distribution, trying to discover differences in class likelihoods.

4. Experiments and Results

This section reports the experiments of the Transformer-based hybrid-fusion model on the speeches in the RAVDESS [12]. It first introduces the dataset formality and feature extraction process (4.1. & 4.2.), then the training and validation stages (4.3. & 4.4.), and the experimental results and their comparisons with State-of-the-art (SOTA) method (4.5.). The section ends with a comparison with other SOTA methods on the RAVDESS dataset (4.6.).

4.1. Dataset

The speeches in the RAVDESS dataset consist of 1440 audio-visual pieces supplied by 12 male and 12 female professional performers. Notably, the lexical contents do not indicate emotional inclination. Each media piece is marked with one of eight emotions: neutral, calm, happy, sad, angry, fearful, disgust, and surprised [12]. Besides neutral, all emotions come at one of two intensities (normal and strong). The speeches in this dataset can be concisely described as each participant reading one of two semantically unemotional sentences and providing two samples for every emotion-intensity-sentence pair [12].

This article has chosen the RAVDESS dataset for experiments, considering three perspectives demonstrating its competence in examining bimodal SER architecture. First, it underwent adequate validation on various dimensions to confirm emotional validity [12]. Second, RAVDESS has been one of the most widely used datasets in SER tasks for its completeness [1]. Third, it is of consistent and well-arranged formality. The speech videos in RAVDESS are of similar length (roughly 3 seconds long), and each is an explicitly labelled utterance [12]. This implies less cumbersome preprocessing to transform data into identical input formats.

4.2. Feature Extraction

Feature extraction of this paper's experiments depends on two open-source technologies, openSMILE and OpenFace [13, 14]. openSMILE is a complete audio-processing toolkit, while OpenFace is a state-of-the-art toolkit for facial behaviour analysis [13, 14].

Visual feature extraction takes advantage of OpenFace's full features: facial landmark detection and tracking, eye gaze tracking, and facial action unit detection [14-18]. 714 features were extracted per frame from the original videos. The videos capture 30 frames per second, which results in 90 to 160 frames per video.

Audio feature extraction uses two parameters, window size and stride length, to create partitions for the program to focus on and extract features. The experiments set a window of 64000 samples while applying dynamical stride lengths to fit the video frames. Thus, each video yields acoustic and visual features with the same sequence length. This allows the model to fuse tensors by simple concatenation.

Eventually, 714 visual features and 6373 audio features were filtered. This excludes any feature which is either temporally constant or repeating another feature for at least 720 samples. As a result, 704 visual features and 1931 audio features stood out. These features were then normalized, and all samples were padded to a length of 160 with their own starting slices. This ensures equal-sized input

for the Transformers. After preliminary filtering, random tests on base Transformer encoders were conducted to select the final 512 features. This saves training and validation time as feature numbers equal to a multiple of the number of GPU processors betters parallel computing.

4.3. Model Training

The implemented model is composed of five modalities: acoustic, visual, intermediate fusion, early fusion, and late fusion, each function as described in Section 3.1. The model benefits from parallel computing as the acoustic, visual, and early fusion modalities could be trained simultaneously. Intermediate fusion modality is trained after acoustic and visual modalities before training late fusion modality to synthesize the four predictions.

The experiment applies stochastic gradient descent (SGD) training on batch sizes of 32 and utilizes the early stopping regularization method with the patience of 10 to decide the number of epochs. Additionally, another configuration is deployed to acquire stable outcomes: Early stopping does not trace validation losses in the first five epochs and does not lose patience in the first ten epochs. This effectively prevented premature stopping in practice as the models do not appear to sufficiently fit the data early on, while a coincidentally low validation loss could terminate training.

4.4. Model Validation

The experiment tunes two hyperparameters, learning rates and L2 regularization weights, through 9-fold cross-validation. Results are respectively demonstrated in Table 1 and Table 2. Each modality has come across at least one set (eight times) of 9-fold cross-validation (72 random tests) for every validated hyperparameter value to expand confidence intervals. The optimal learning rate for five component modalities are 5e-4, 1e-4, 2.5e-5, 5e-5, and 2.5e-5; while the optimal L2 regularization weights are 2e-6, 5e-7, 0, 2e-6, and 1e-7.

Table 1. Accuracies on different learning rates for late fusion, with no regularization

Learning Rate	1.25e-5	2.5e-5	5e-5	1e-4	2e-4
Accuracy	90.56%	90.7%	90.69%	90.13%	89.08%

Table 2. Accuracies on different L2 regularization weights for late fusion, at learning rate 2.5e-5

L2	0	1e-7	5e-7	1e-6	5e-6	1e-5
Accuracy	90.7%	91.5%	91.15%	91.04%	90.88%	90.78%

4.5. Statistical Results

The experimental results can be summarized in a confusion matrix, as displayed in Fig. 2. The average performance of different fusion strategies and loss functions are demonstrated in Table 3 and Table 4, respectively. The average accuracy in a set of cross-validation was 91.1%. For all eight classes, the model achieves recalls above 85%. The results also demonstrate enhanced capabilities with the proposed DCE loss function. Furthermore, DCE loss collaborates well with the early stopping method. Within a set of cross-validations using the CE loss function on the acoustic modality, the largest observed difference between the highest validation accuracy in all training epochs and the one adopted by early stopping hits nearly 20%. In contrast, such differences stay under 10% with DCE loss.

Table 3. Average modality accuracies on at least eight 9-fold cross-validation

Learning Rate	1.25e-5	2.5e-5	5e-5	1e-4	2e-4
Accuracy	90.56%	90.7%	90.69%	90.13%	89.08%

Table 4. Accuracies with CE and DCE loss functions

Modality	Intermediate Fusion	Early Fusion	Late Fusion
Acc w/ CE	86.1%	88.6%	89.7%
Acc w/ DCE	87.8%	90.2%	91.3%

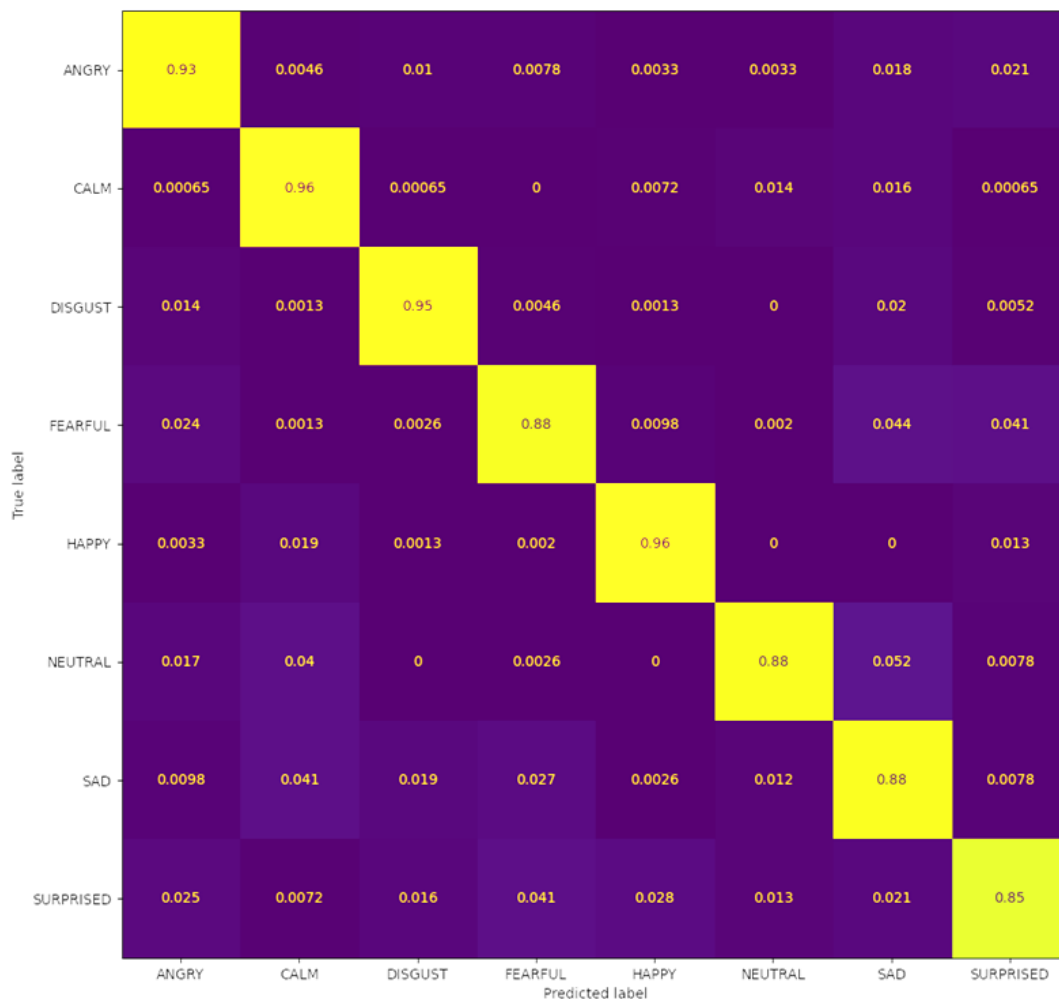


Figure 2. Confusion matrix on a set of 9-fold cross-validation (Figure Credits: Original)

4.6. Comparison with SOTA Methods

Among the presented SOTA methods in Table 5, the first three models were unimodal SER using Transformers [19-21]. Moreover, the following three works were multimodal SER without the use of Transformers [22-24] and two works utilized the attention mechanism [23, 24].

Compared with these recent studies published after 2023, the proposed method stands out not only with a better statistical result but also with the result reliability thanks to sufficient cross-validation. The joint fusion architecture, for example, published their best accuracy [22]. In comparison, the results in this article take averages of at least eight 9-fold cross-validations, minimizing biases from different dataset splits.

Table 5. Comparison with SOTA methods

Method	Accuracy
SCQT-MaxViT [19]	77.54%
Mel-MViTv2 [20]	81.75%
MaxMViT-MLP [21]	89.12%
Joint Fusion Architecture [22]	89%
LSTM + Attention + CNN-2D [23]	74.44%
AT + VS + VC + VT [24]	89.25%
Proposed Transformer-based Hybrid Fusion Method	91.1%

5. Discussion

On average, the architecture achieves 91.1% accuracy. It exhibits outstanding aptitudes for identifying calm and happy, with recalls of 96%. The predictive ability is insensitive to hyperparameters since the variation in both the learning rates and the L2 regularization weights only slightly disturbs the performance.

Multi-modal design proves to outperform single modality: both the intermediate and the early fusion attain higher accuracies at (88.2% and 89.8%). The late fusion modality, in particular, climbs to 91.1% accuracy with only 32 extra parameters. Each of these 32 weights simulates a percentage credibility of the corresponding modality on the corresponding class. Accordingly, it promotes intramodal strengths and offsets weaknesses.

The proposed DCE loss exerts remarkable positive effects on model capabilities. It not only raises accuracies but also increases model stability. Its consideration for degrees of incorrectness, on top of the uncertainty in prediction (as the classic CE loss focuses on), benefits the Transformer-based SER tasks involved in this study.

On the downside, it left space for improvement in predicting surprised emotion (recall at 85%). The surprised emotion is easily confused with fearful (4.1%), happy (2.8%), and angry (2.5%). Similarly, samples of these emotions are likely to be misclassified as surprised samples (4.1%, 1.3%, and 2.1%). This may indicate an emotional association in the dataset.

The outperformance of the proposed method reemphasizes the Transformers' suitability for SER tasks and verifies the advantages of hybrid fusion on emotion classification.

6. Conclusion

This paper elaborates a Transformer-based hybrid fusion approach for speech emotion recognition. This work discusses the theoretical backgrounds and practical experimentation with the architecture. This architecture relies on a late fusion modality that assesses the outcomes of other unimodal and multimodal predictors. Moreover, the proposed DCE loss function deriving from the CE loss and its influence over model behaviors are also discussed. Finally, the proposed design with other SOTA methods is compared. For future works, this architecture can be tested on other datasets to confirm its general applicability. More modalities, such as text or physiological signals, can be employed with more advanced fusion methods than concatenation. Besides, as this study only considers transformer encoders, it is recommended that the feasibility of combining transformer decoders with the current design be explored.

References

- [1] De Lope, Javier, Manuel Graña. An ongoing review of speech emotion recognition. *Neurocomputing*, 2023, 528: 1-11.
- [2] Gandhi Ankita, Adhvaryu Kinjal, Poria Soujanya, et al. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 2023, 91: 424-444.
- [3] Khare Smith K, Blanes-Vidal Victoria, Nadimi Esmaeil S, et al. Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Information Fusion*, 2023: 102019.
- [4] Vaswani Ashish, Shazeer Noam, Parmar, Niki, et al. Attention is all you need. *International Conference on Neural Information Processing Systems*, 2017: 6000-6010.
- [5] Wagner Johannes, Triantafyllopoulos Andreas, Wierstorf Hagen, et al. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45 (9): 10745-10759.
- [6] Xu Jie, Huang Feiran, Zhang Xiaoming, et al. Sentiment analysis of social images via hierarchical deep fusion of content and links. *Applied Soft Computing*, 2019, 80: 387-399.

- [7] Aziz Abdul, Chowdhury Nihad Karim, Kabir Muhammad Ashad, et al. MMTF-DES: A fusion of multimodal transformer models for desire, emotion, and sentiment analysis of social media data. arXiv preprint, 2023: 2310.14143.
- [8] Xie Jie, Mingying Zhu, Kai Hu. Fusion-based speech emotion classification using two-stage feature selection. *Speech Communication*, 2023, 152: 102955.
- [9] Gan Chenquan, Fu Xiang, Feng Qingdong, et al. A multimodal fusion network with attention mechanisms for visual–textual sentiment analysis. *Expert Systems with Applications*, 2024, 242: 122731.
- [10] Bahdanau Dzmitry, Kyunghyun Cho, Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint, 2014: 1409.0473.
- [11] Ozsoy Makbule Gulcin. Multi-Margin Loss: Proposal and Application in Recommender Systems. arXiv preprint, 2024: 2405.04614.
- [12] Livingstone Steven R., Frank A. Russo. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one*, 2018, 13 (5): e0196391.
- [13] Eyben Florian, Martin Wöllmer, Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM international conference on Multimedia*. 2010: 1459-1462.
- [14] Baltrušaitis Tadas, Peter Robinson, Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. *IEEE winter conference on applications of computer vision*, 2016: 1-10.
- [15] Zadeh Amir, Chong Lim Yao, Baltrusaitis Tadas, et al. Convolutional experts constrained local model for 3d facial landmark detection. *Proceedings of the IEEE international conference on computer vision workshops*. 2017: 2519-2528.
- [16] Baltrusaitis Tadas, Peter Robinson, Louis-Philippe Morency. Constrained local neural fields for robust facial landmark detection in the wild. *Proceedings of the IEEE international conference on computer vision workshops*. 2013: 354-361.
- [17] Wood Erroll, Baltrusaitis Tadas, Zhang Xucong, et al. Rendering of eyes for eye-shape registration and gaze estimation. *Proceedings of the IEEE international conference on computer vision*. 2015: 3756-3764.
- [18] Baltrušaitis Tadas, Marwa Mahmoud, Peter Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. *IEEE international conference and workshops on automatic face and gesture recognition*. 2015, 6: 1-6.
- [19] Mocanu Bogdan, Ruxandra Tapu, Titus Zaharia. Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning. *Image and Vision Computing*, 2023, 133: 104676.
- [20] Ong Kah Liang, Lee Chin Poo, Lim Heng Siong, et al. SCQT-MaxViT: Speech Emotion Recognition with Constant-Q Transform and Multi-Axis Vision Transformer. *IEEE Access*, 2023, 11: 63081-63091.
- [21] Ong Kah Liang, Lee Chin Poo, Lim Heng Siong, et al. Mel-MViTv2: Enhanced speech emotion recognition with mel spectrogram and improved multiscale vision transformers. *IEEE Access*, 2023, 11: 108571-108579.
- [22] Haddad Syrine, Olfa Daassi, Safya Belghith. Single Modality and Joint Fusion for Emotion Recognition on RAVDESS Dataset. *SN Computer Science*, 2024, 5 (6): 669.
- [23] Singh Jagjeet, Lakshmi Babu Saheer, Oliver Faust. Speech emotion recognition using attention model. *International Journal of Environmental Research and Public Health*, 2023, 20 (6): 5140.
- [24] Ong Kah Liang, Lee Chin Poo, Lim Heng Siong, et al. MaxMViT-MLP: Multiaxis and Multiscale Vision Transformers Fusion Network for Speech Emotion Recognition. *IEEE Access*, 2024, 12: 18237-18250.