Research And Future Application Analysis of Multimodal Fusion

Haowen Xue [1,*], Zhitao Zhu [2]

[1] Chongqing University of Posts and Telecommunications, Chongqing, China

[2] Beijing University of Posts and Telecommunications, Beijing, China

* Corresponding author: 2021214872@stu.cqupt.edu.cn

Abstract. This manuscript delves into the origin, progression, and future of multimodal fusion by conducting a comprehensive review of seminal research at various phases of its ontogeny. It encompasses three foundational methodologies of multimodal integration: Early Fusion, late Fusion, and Hybrid Fusion. Moreover, the article presents three novel methodologies for integration. It culminates in a discourse on potential subsequent applications, obstacles, and progress within the domain. solve the fusion problem of heterogeneous neural networks, and improve the correlation and consistency between modes in the aspects of cross-modal representation learning, multi-modal sentiment analysis, multi-modal intelligent interaction, so as to achieve better human-computer interaction experience.

Keywords: Multimodal fusion, machine learning, early fusion, late fusion, hybrid fusion sentiment analysis.

1. Introduction

Each source or form of information can be referred to as a modality. For example, humans have touch, hearing, vision, and smell; information media include voice, video, and text. In the general sense, multimodal learning refers to using models to simultaneously process multiple types of data, such as handling image-text combinations, generating text from images, or generating images from text. By using large multimodal models, we can better understand and process complex multimodal data, improving the performance of AI applications.

Multimodal fusion refers to the integration and joint processing of different sensory modalities (such as vision, hearing, and speech) to obtain richer and more comprehensive information.

Initial Research Phase (before 2000): During this stage, the concept of multimodal fusion was just emerging. Experts and scholars began exploring how to integrate different sensory modalities to improve the performance of intelligent systems.

Feature Fusion Phase (2000 to 2010): As research progressed, scholars started focusing on how to fuse the features of different sensory modalities.

Model Fusion Phase (2010 to present): With advances in technology, researchers began exploring how to integrate deep learning models of different sensory modalities and achieve joint learning of multimodal information through methods like shared and cross-training.

Heterogeneous Fusion Phase (future trend): Current research mainly focuses on heterogeneous data fusion, which involves integrating data from different domains and representations.

This paper mainly aims to outline several different fusion methods, exploring their advantages and disadvantages, to help readers quickly understand multimodal fusion and make predictions for future developments.

2. Typical Theories and Methods for Multimodal Fusion

2.1. Early Fusion

Early fusion is also known as feature fusion. The basic idea is that after feature extraction of the input modalities, these features are merged into one fusion feature and this fusion feature is fed into

the model to finally obtain the recognition results (can be seen in Fig. 1). Early fusion is able to capture low-level correlation information between different modalities and can be combined with feature extraction methods to weed out redundant information. For example, mRMR, Autoencoders and other techniques are used to reduce the dimensionality of the features and eliminate the redundancy problem in the input. However, the fusion features produced by transforming and scaling the modal features in the early fusion usually have high dimensionality, and the fusion features can be dimensionality-reduced using principal component analysis (PCA) and linear discriminant analysis (LDA).

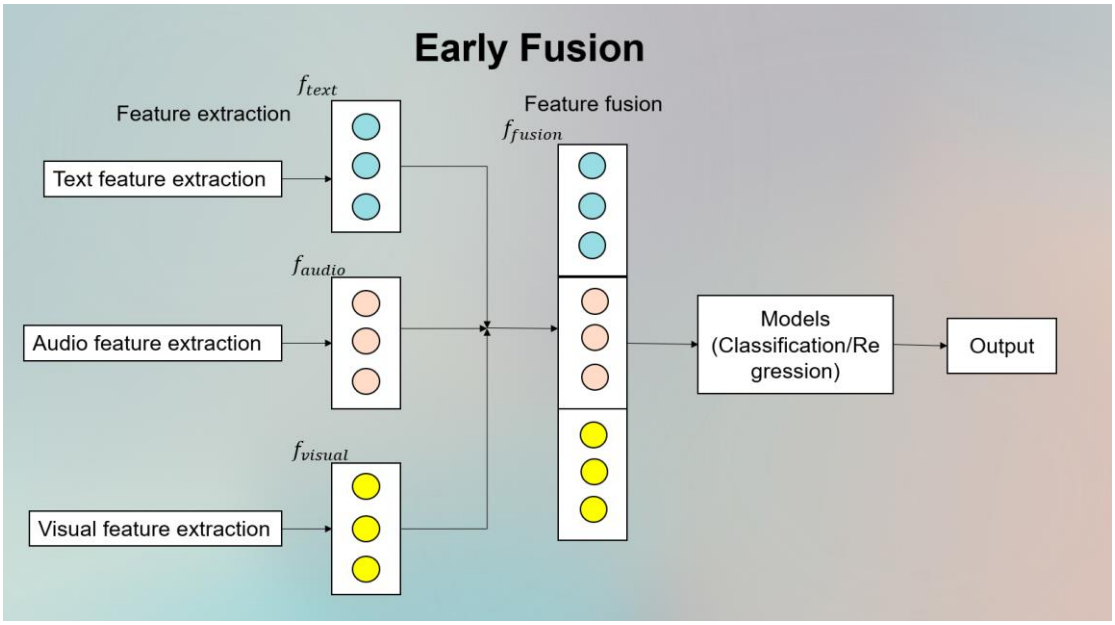

Figure 1. Early Fusion

In early fusion, how to accurately capture hidden time dependencies in the data is a difficult problem to overcome. It is essential for understanding changes in emotional states in audio signals. The multi-scale temporal dilated CNN (MS-TDCNN) has posed one method to solve this problem. [1]. In this method, text feature extraction utilizes the Stanford Sentiment Treebank (SST) dataset, and further employs the BERT model for in-depth contextual text feature extraction. For audio feature extraction, the AVEC dataset is used, and spectral and power spectral features are extracted using the openSMILE toolkit. These features are segmented across multiple time spans, and various statistical functions are applied to extract primary characteristics from the audio signals.

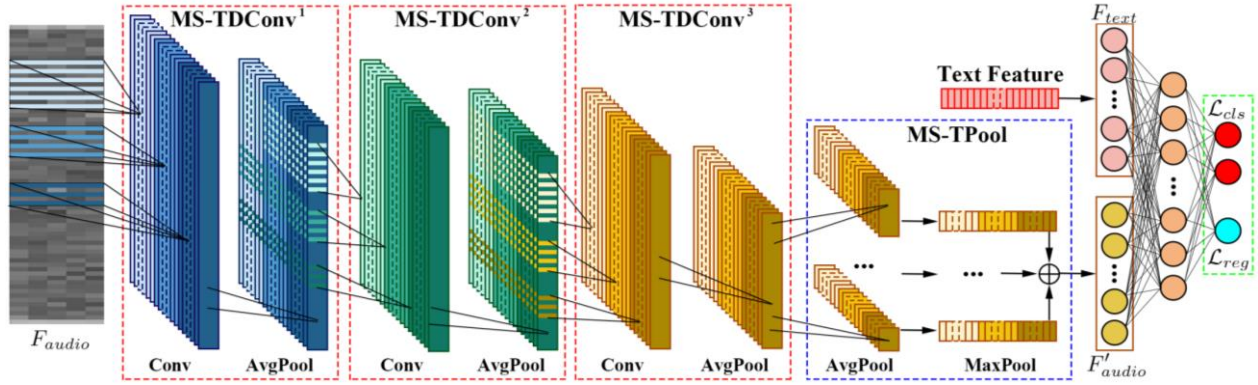

Figure 2. The architecture of the multi-scale temporal dilated CNN (MS-TDCNN)

In order to more accurately capture the changes in pitch in speech signals, the method proposes a multi-scale time temporal dilated CNN(MS-TDCNN) architecture [1] (The architecture of MS-TDCNN can be in Fig. 2). The architecture is divided into two main parts: multi-scale time dilation convolution (MS-TDConv) and multi-scale time pooling (MS-TPool) [1]. MS-TDConv increases the

receptive field of the convolution kernel by inserting intervals between its elements to more accurately capture affective states in the audio signal. MS-TPool is used to improve generalization and avoid overfitting. Finally, the text feature vector and audio feature vector are spliced before the fully connected layer as input to the subsequent detection model.

The experimental results show in Table 1 that the method performs better in terms of RMSE and MAE and has a higher CCC than the traditional method. The method captures the hidden time dependencies in the data well. The method has the potential to enable real-time or near real-time depression detection, providing rapid feedback and intervention for patients. It can also be integrated into smart devices or apps for continuous monitoring of the user's mental state and providing alerts or suggestions when necessary.

Table 1. Performance comparison between different models

Model	Dev			Test		
	RMSE	MAE	CCC	RMSE	MAE	CCC
Official Baseline (Audio)	6.43	-	.272	8.19	-	.045
Official Baseline (Video)	7.72	-	.269	8.01	-	.120
Official Baseline (Fusion)	5.03	-	.336	6.37	-	.111
Baseline	6.20	4.88	.348	6.80	5.34	.255
Baseline + Multi-modality + Multi-task	5.62	4.44	.397	6.64	5.10	.281
Baseline + Multi-modality + Multi-task + MS-TDConv +MS-TPool	5.12	3.89	.463	6.35	4.93	.385
Model Ensemble (Final)	5.07	4.06	.466	5.91	4.39	.430

2.2. Late Fusion

Late Fusion, also known as Decision Fusion, means that each modal data is processed by an independent model and generates initial classification results, and then these independent results are fused to produce a more reliable decision (The architecture of late fusion can be seen in Fig. 3). Compared to early fusion, late fusion can handle the asynchronous nature of the data more simply. The whole system can be scaled up as the number of modalities increases, and the exclusive prediction model for each modality can be better modelled for that modality, and can also make predictions when the model inputs are missing for certain modalities. However, compared to early fusion, late fusion does not take into account modal correlations at the feature level and is more complex and difficult to implement.

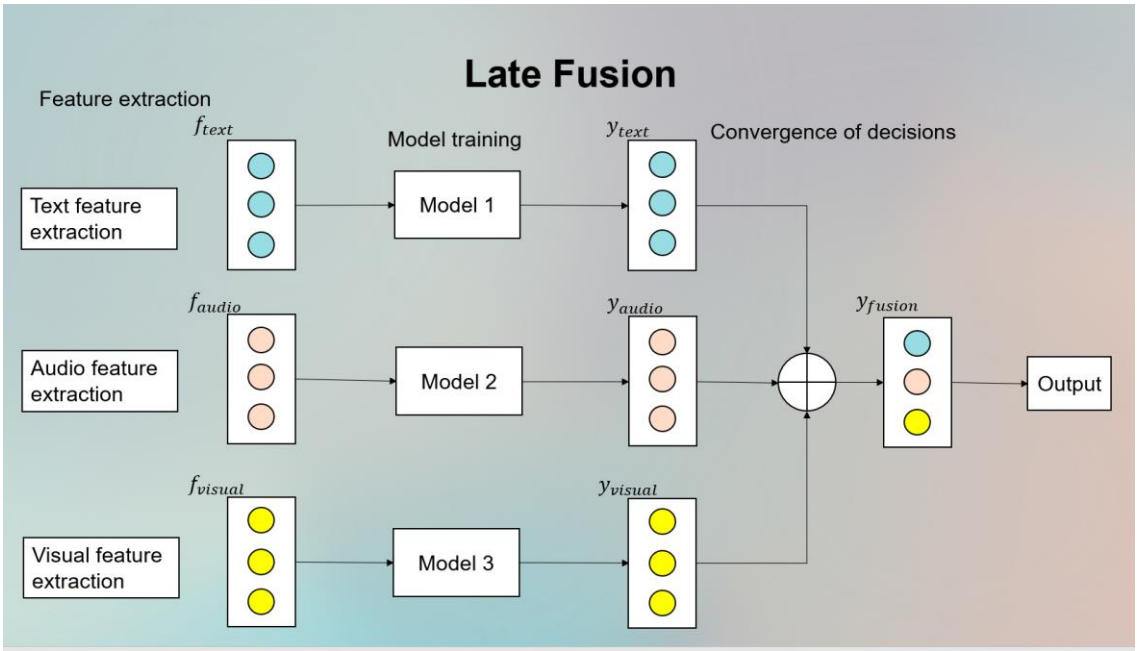

Figure 3. Late Fusion

Traditional models of emotion perception in the temporal dimension (e.g., the MS-TDCNN model) are only applicable to unimodal modalities, and do not take into account the short-term continuity of emotion change in multimodal contexts. Yishan Chen, Zhiyang Jia et al. proposed a Context-Aware Decision-Level Fusion (CADLF) model for audiovisual modality-based MEP-AV architecture [2]. The model proposes an architecture that contains four modules: A data acquisition module, Audio Expression Analysis Module, Visual Expression Analysis Module, and a Multimodal Fusion Module. (The architecture of MEP-AV can be seen in Fig. 4)

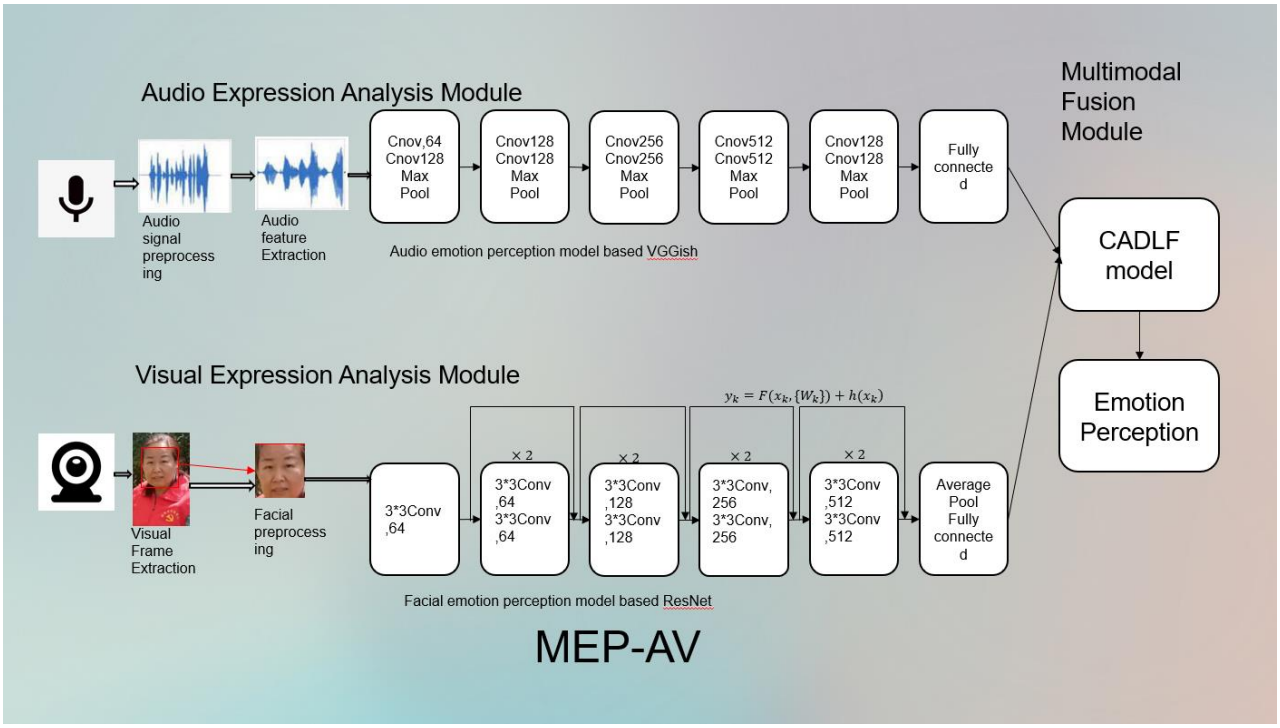

Figure 4. The architecture of MEP-AV

The IEMOCAP dataset used in the data acquisition module contains improvisations by two speakers to record expressive videos. It contains approximately 12 hours of audiovisual data, including videos, speeches, and text transcriptions. Each utterance in the dialogue box is labelled by discrete and continuous emotion tags [2].

The Audio Expression Analysis Module contains three parts: Audio signal preprocessing, Audio feature extraction and Audio emotion perception. The main part of Audio emotion perception model is VGGish model (The architecture of it can be seen in Fig. 4). This model is used to get the emotion perception of audio modality.

The Visual Expression Analysis Module also contains three parts: Visual Frame Extraction, Facial preprocessing and Facial emotion perception. In Facial emotion perception model, the main part of Facial emotion perception model is ResNet model (The architecture of it can be seen in Fig. 4). This model is used to obtain a facial emotion perception

The multimodal fusion module uses decision fusion models to obtain emotion perception results. Previous research on decision-level fusion models has relied on traditional machine learning methods. These methods have performed poorly in terms of short-term continuity of each model [3-5]. Addressing the temporal correlation of audio and facial emotion, the proposed CADLF model utilises contextual information of multimodal sentiment to estimate the emotion in the current state to satisfy short-term continuity. In order to create links between the various modal and temporal dimensions, Long Short-Term Memory Network (LSTM)nodes were used to construct the CADLF model. LSTM networks are recurrent neural networks capable of learning long-term dependencies in time series without encountering gradient vanishing or explosion [6]. Despite the superior performance of LSTM in emotion prediction, single-layer LSTM networks still face the problem of not being able to handle complex data. In order to overcome the short-term continuity of the output, results mentioned earlier,

a three-layer stacked LSTM was used for sentiment information fusion (The architecture of CADLF can be seen in Fig. 5). The vertical structure creates a hierarchical feature representation and improves the processing of complex input data.

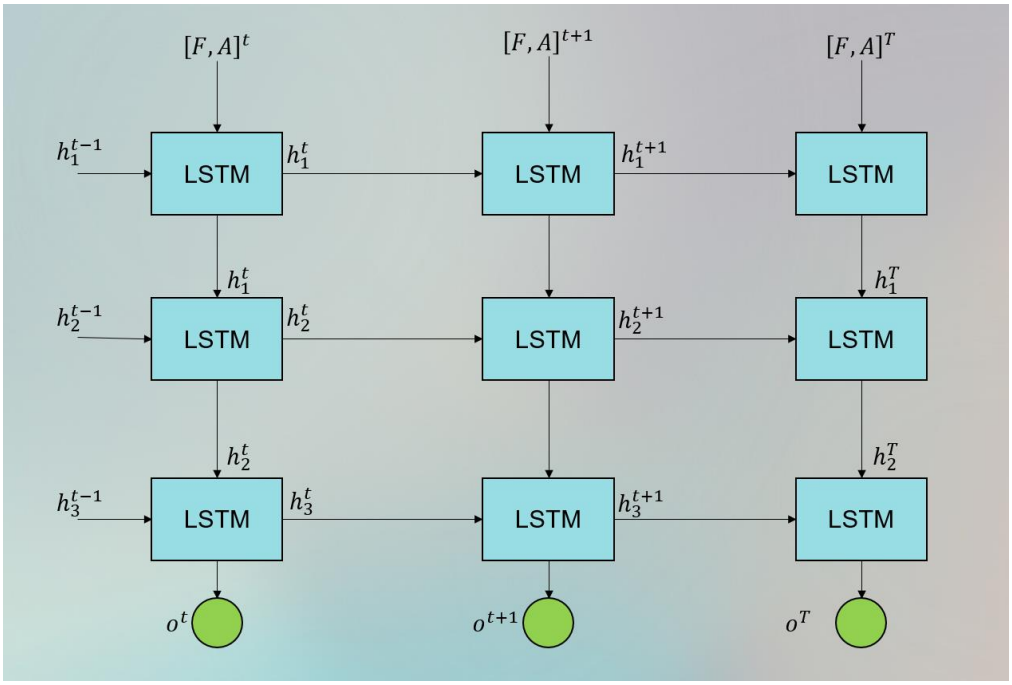

Figure 5. The architecture of CADLF

The experimental results show in Table 2 that this model significantly improves the accuracy and F1-score of continuous sentiment prediction, the RMSE is reduced to less than 1, and the PCC (Pearson's correlation coefficient) is significantly better than other models. The short-term continuity of emotion change in multimodality is well addressed.

Table 2. Performance comparison of MEP-AV model with unimodal model

Model	Accuracy	F1-score	RMSE	PCC
Audio emotion Perception model	38.03%	26%	2.059	0.273
Facial emotion perception model	68.56%	68%	1.082	0.841
MEP-AV model	70.89%	70.2%	0.804	0.904

2.3. Hybrid Fusion

Hybrid fusion can be performed multiple times at different stages. For example, partial feature fusion can be performed early in the model, followed by further fusion at an intermediate or late layer.(The architecture of Hybrid Fusion can be seen in Fig. 6). In contrast to early and late fusion, hybrid fusion allows models to learn feature interactions and combinations at different levels. It improves the expressive power of the model and also enhances the ability to learn modalities in complex scenarios. However, it should be noted that if the hybrid features are too complex, it may increase the risk of overfitting.

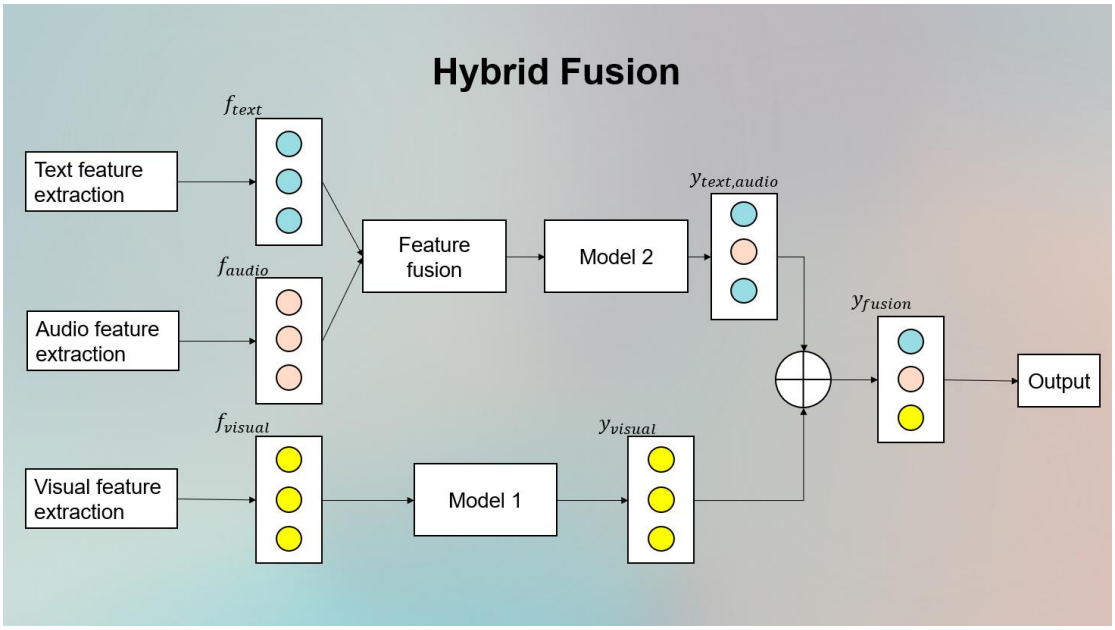

Figure 6. Hybrid Fusion

Traditional unimodal emotion recognition methods (based only on facial expressions or video or one modility) suffer from difficulty in improving accuracy because emotion expressions are complex and may contain artifacts or expressions of opposite meanings, leading to misclassification of emotion recognition.[7] Guang Yang et al. proposed a hybrid multimodal emotion recognition model based on LSTM and optimal weight search algorithm(Fig 7) [7]. The model incorporates not only two non-physiological signals, speech and facial expression, but also a physiological signal, electroencephalography (EEG), to form a hybrid fusion strategy. This strategy captures the complexity of emotional expression in a more comprehensive way. The model proposes an architecture that contains four modules: Data acquisition and processing module, Speech and video sentiment analysis and feature fusion model, Brainwave Feature Extraction and sentiment Analysis module, and Decision fusion module. (The architecture of this model can be seen in Fig. 7)

In the speech and video sentiment analysis and feature fusion model, LFCNN (Lightweight Fully Convolutional Neural Network) and improved GhostNet models are used for feature extraction and feature level fusion by the LSTM model.(The architecture of it can be seen in Fig. 7)

In the brainwave feature extraction and sentiment Analysis module, spatial features are extracted using a CNN network.

The key method in Decision Fusion is the Optimal weight search algorithm, this method have 3 steps: Weight iteration and condition checking, Predict wake-up scores and Calculate the RMSE and update the optimal weights.(Optimal weight search algorithm in Fig. 8)

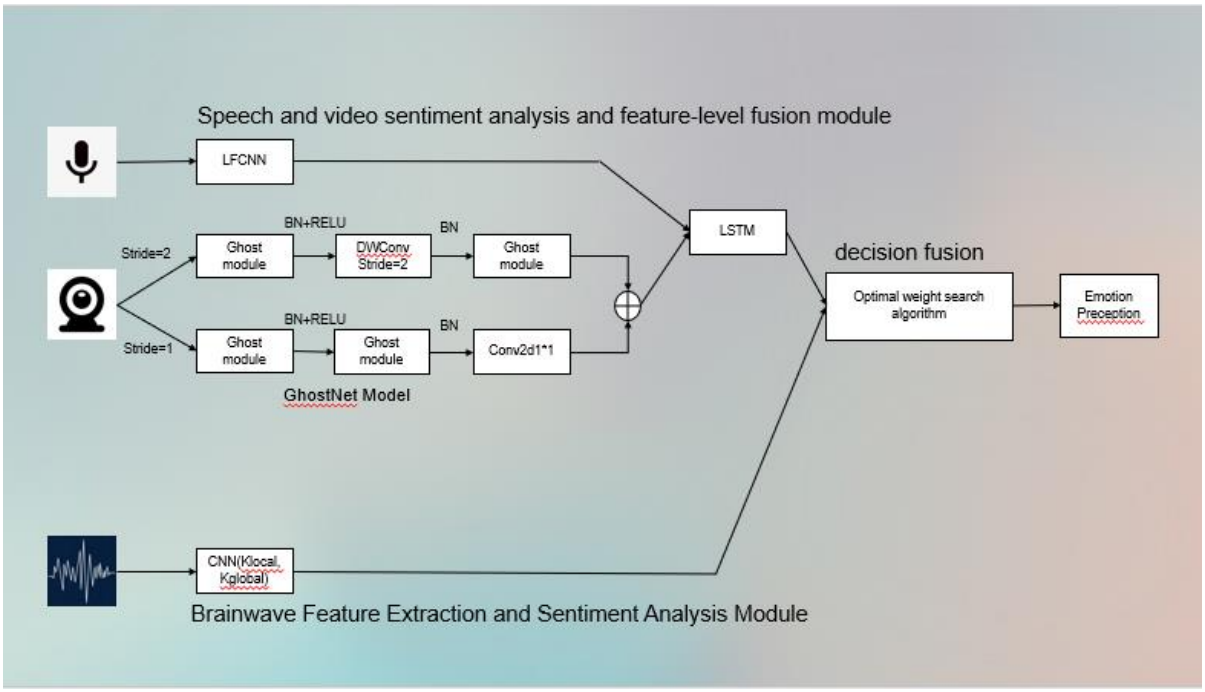

Figure 7. A Hybrid multimodal emotion recognition model based on LSTM and optimal weight search algorithm

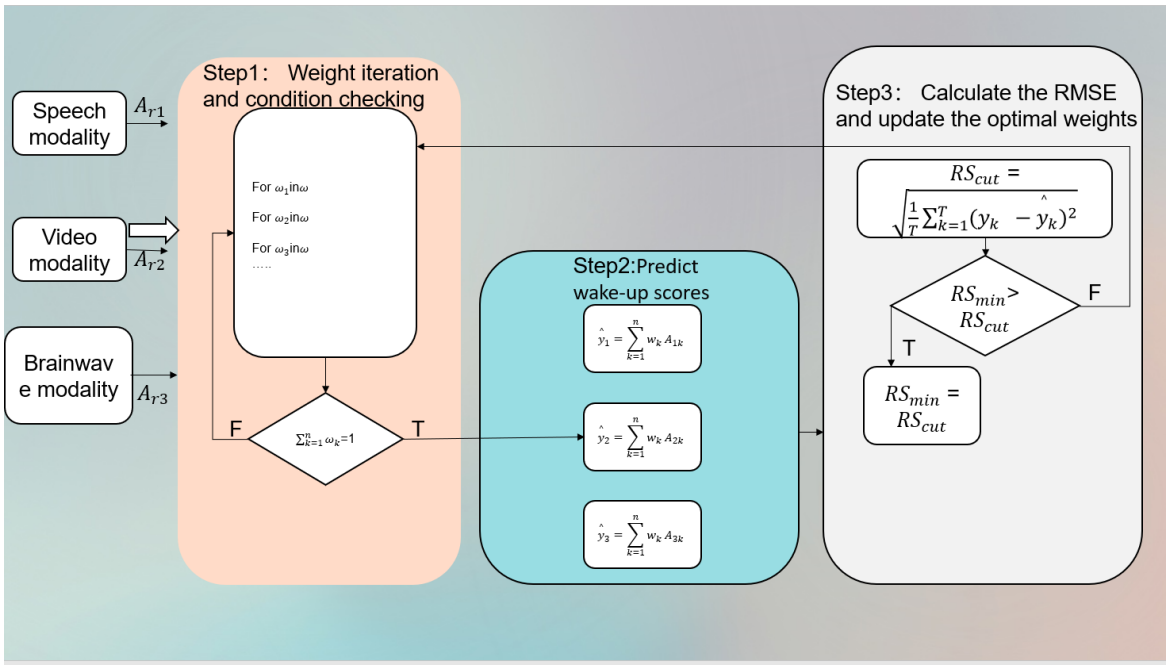

Figure 8. Decision Fusion(based on Optimal weight search algorithm)

The experimental results show that the method has a significant improvement in accuracy and F1-score on the MAHNOB-HCI dataset compared to recent models (Table 3). And when tested on other datasets (Table 4), the accuracy is above 90% and the F1-score is above 85%. The method resulted in a very significant improvement in emotion recognition accuracy by fusing 3 modalities (1 physiological, 2 non-physiological).

Table 3. Model Performance Comparison

Method	Signals	ACC(%)	F1(%)
1D-CNN-24	EEG,GSR,RSP	73.32%	71.74%
ST-GCNBILSTM	Body movement,speech,text	78.74%	76.89%
AM-C3D	Face speech	80.22%	79.39%
SC-DTCN	Face EGG	88.57%	84.32%
MFFNN	EGG EOG	87.32%	85.04%
Our method	EEG speech,face	94.53%	91.26%

Table 4. Performance of the model on different datasets

Data set	ACC(%)	F1(%)
SEED	92.89%	88.62%
Self-speech	91.54%	85.36%

3. Future Applications and Challenges

Modal fusion has the potential to revolutionize intelligent driving and traffic management. By integrating data from various sensors, such as cameras, radar, and lidar, it enables a more comprehensive understanding of the environment, leading to safer and more efficient transportation systems. In virtual reality (VR) and augmented reality (AR), modal fusion can enhance immersive experiences by blending sensory inputs like vision, hearing, and touch, thereby expanding how people perceive and interact with virtual environments. This allows for the creation of more realistic and engaging experiences that push the boundaries of user interaction. For security monitoring, modal fusion improves the effectiveness of surveillance systems by merging data from video, audio, infrared, and other sensors. This multi-dimensional approach enables more accurate target detection, behavior analysis, and timely security alerts, leading to more efficient event management and response.

A key challenge for the future lies in how to effectively fuse heterogeneous neural networks. In deep learning, networks with varying depths typically possess different representational and learning capabilities. By merging these networks, we can potentially achieve better performance and results. The applications of multimodal fusion are extensive. Cross-modal representation learning is one area where future research could focus on developing more effective methods to improve the correlation and consistency between different modalities. This would enhance the integration of information across various sensory inputs. Cross-modal information fusion is another promising avenue, where future work could develop advanced algorithms and models to combine modalities like image and text, or audio and text. Such research could lead to more seamless and comprehensive understanding of diverse data sources. Multimodal intelligent interaction is a third area of interest, where future innovations could focus on technologies that integrate multiple modes of interaction, such as voice, gestures, and facial expressions. This would create more natural, intuitive, and efficient interactive experiences for users.

4. Conclusion

This paper primarily discusses the development of multimodal fusion and introduces three fundamental theories: early fusion, late fusion, and hierarchical integration. Additionally, it presents a novel approach to three new fusion modes. The paper also provides three real-life applications where multimodal fusion enhances performance.

Future research will focus on heterogeneous fusion to achieve a full integration of neural networks at different depths. Multimodal fusion will also be applied to a broader range of scenarios, further promoting AI development across various fields.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Fan Weiquan, He Zhiwei, Xing Xiaofen, Cai Bolun, Lu. Weirui Multi-modality Depression Detection via Multi-scale Temporal Dilated CNNs Session: Detecting Depression with AI Sub-challenge, 2019.

[2] Yishan Chen, Zhiyang Jia, Kaoru Hirota, Yaping Dai. P A Multimodal Emotion Perception Model based on Context-Aware Decision-Level Fusion, roceedings of the 41st Chinese Control Conference, 2022.

[3] J. Ye, W. Zheng, and L. Yang, et al, Multimodal emotion recognition based on deep neural network, Journal of Southeast University (English Edition), 2017, 33 (4): 444 - 447.

[4] F. Xu and Z. Wang, Emotion Recognition Research Based on Integration of Facial Expression and Voice, 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2018: 1 - 6.

[5] Q. Xu, B. Sun, and J. He, et al, Multimodal Facial Expression Recognition Based on Dempster-Shafer Theory Fusion Strategy, 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), 2018: 1 - 5.

[6] F. Karim, S. Majumdar, H. Darabi and S. Chen, LSTM Fully Convolutional Networks for Time Series Classification, in IEEE Access, 2018, 6: 1662 - 1669.

[7] Guang Yang Jiaxin Li Deguo Yang, Guojun Wang 2023 5th International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI) Multimodal Emotion Recognition Based on Hybrid Fusion, 2023.