



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642

<https://www.inderscience.com/ijict>

Multimodal emotion feature extraction and information fusion methods for video content

Jianxiao Ma

DOI: [10.1504/IJICT.2026.10077026](https://doi.org/10.1504/IJICT.2026.10077026)

Article History:

Received:	26 September 2025
Last revised:	20 October 2025
Accepted:	22 October 2025
Published online:	26 March 2026

Multimodal emotion feature extraction and information fusion methods for video content

Jianxiao Ma

School of Intelligent Manufacturing and Electrical Engineering,
Nanyang Normal University,
Nanyang, 473061, China
and
Collaborative Innovation Center of Intelligent
Explosion-Proof Equipment,
Henan Province, Henan, 473000, China
Email: majianxiao198511@163.com

Abstract: Multimodal sentiment analysis remains challenging due to the difficulty of fusing heterogeneous data like facial expressions, speech, and pose. Unlike unimodal analysis, it is often hindered by problems of information redundancy, heterogeneity, and complex temporal dynamics. This paper proposes MERCAT, a Transformer-based model featuring a cross-modal self-attention mechanism to capture deep correlations between different modalities. This design enables highly efficient and context-aware inter-modal fusion. Extensive experiments on multiple benchmarks show that MERCAT achieves excellent performance. It notably excels in emotion classification, significantly improving accuracy and F1-score over strong baselines, and in emotion intensity prediction, where it substantially reduces error and improves correlation. The study conclusively verifies the efficacy of the cross-modal attention mechanism for information fusion, providing a robust and effective solution for advancing multimodal sentiment analysis.

Keywords: multimodal sentiment analysis; sentiment feature extraction; information fusion; transformer architecture; cross-modal self-attention mechanism; sentiment intensity prediction.

Reference to this paper should be made as follows: Ma, J. (2026) 'Multimodal emotion feature extraction and information fusion methods for video content', *Int. J. Information and Communication Technology*, Vol. 27, No. 28, pp.43–59.

Biographical notes: Jianxiao Ma attended Beihang University from 2004 to 2008, majoring in Detection, Guiding and Control Technology for his undergraduate studies. From 2008 to 2011, he completed his Master's degree at the PLA University of Science and Technology, specialising in Systems Engineering. He is currently a Lecturer at Nanyang Normal University. His main research areas include intelligent control and virtual simulation. In recent years, he has led and participated in five provincial key scientific research projects, obtained eight utility model patents and defence invention patents, and published more than ten academic papers in domestic and international journals.

1 Introduction

Sentiment analysis is a critical research area in artificial intelligence, with much of the focus traditionally on text, speech, and images. However, analysing sentiment in video content presents unique challenges due to its multimodal nature, including the need to fuse information from facial expressions, audio, and body posture. Existing methods often struggle with issues such as inter-modal heterogeneity, temporal variability, and information redundancy (Wang et al., 2025b), leading to limited performance in real-world scenarios. To address these challenges, we introduce the multimodal emotion recognition with cross-attention transformers (MERCAT) model. MERCAT integrates a transformer architecture with a cross-modal self-attention mechanism, enabling more effective fusion of multimodal features from video content (Wang et al., 2025a). This approach significantly improves sentiment classification accuracy and emotion intensity prediction, offering a more robust solution to the problem of multimodal sentiment analysis.

With the rapid development of technologies such as virtual reality (VR), augmented reality (AR), intelligent customer service, human-computer interaction and video recommendation, the application scenarios of sentiment analysis are increasing (Li and Guenier, 2024; DeMatteo et al., 2024). Sentiment analysis can not only enhance the user experience and help virtual assistants better understand the emotional needs of users, but also play an important role in intelligent driving, health care and other fields. For example, by recognising the driver's emotional state (Xing et al., 2025; Ning et al., 2025), the car can determine whether there is a risk of fatigue driving (Xing et al., 2019); in the field of health care, sentiment analysis can help monitor the emotional changes of the elderly or patients, and timely detection of mental health problems (Cao et al., 2025; Li et al., 2025).

Although multimodal sentiment analysis has great potential for application in video content, it still faces many challenges. First, the information of different modalities (e.g., facial expression, speech, and pose) is heterogeneous (Huang et al., 2025; Meng and Lin, 2023), and it is a major challenge to effectively fuse this information to ensure that the model is able to understand the inter-modal relationships. Second, sentiment analysis is not only limited to the categorisation of sentiment categories (Qi et al., 2022; Huang, 2025), but also involves the estimation of sentiment intensity, and different individuals have different emotional responses (Zeng, 2023), which puts higher requirements on the model. In addition, the temporal characteristics of video content also increase the complexity of sentiment analysis (Abinaya and Kumar, 2024; Poloju and Rajaram, 2024), which needs to take into account the dynamic changes of sentiment expression.

In order to solve the above problems, this study proposes a multimodal emotion feature extraction and information fusion method based on video content. By combining information from multiple modalities, such as facial expression, audio, and pose, an innovative multimodal sentiment analysis framework is designed to more accurately identify the sentiment categories in the video and estimate the intensity of the sentiment. The MERCAT model integrates a transformer architecture with a cross-modal self-attention mechanism, enabling more effective fusion of multimodal features from video content. Cross-modal self-attention refers to a mechanism that allows the model to compute relationships between features from different modalities, such as facial expressions, audio, and body posture. Unlike traditional attention mechanisms that focus on a single modality, cross-modal attention helps the model dynamically adjust the

importance of different modalities by learning how they interact with each other. This is particularly useful in multimodal sentiment analysis, where different types of data (e.g., facial expressions and audio) may convey complementary emotional cues.

The primary contributions of this research are outlined as follows:

- 1 introduce a feature fusion-based framework for multimodal sentiment analysis, capable of effectively combining data from vision, audio and pose modalities
- 2 the transformer architecture and attention mechanism are introduced to deeply fuse cross-modal information and enhance the precision of sentiment analysis
- 3 in the task of sentiment intensity prediction, a regression model is designed that can accurately predict the intensity of sentiment.

This study not only provides an innovative multimodal model for the field of sentiment analysis, but also provides theoretical basis and technical support for the task of sentiment recognition in practical applications.

2 Research design

2.1 Holistic modelling: framework, design, and network architecture

In the current field of multimodal sentiment analysis, although a number of studies have adopted deep learning techniques to enhance the effectiveness of sentiment analysis, they still face many challenges. First, sentiment expressions are diverse and heterogeneous across modalities, which make inter-modal information fusion very difficult. Second, emotional information in video content is often temporal in nature (Jawahar et al., 2021), and it is also a challenge to effectively capture emotional changes between video frames. Although there have been some models that use fusion methods for facial expression, speech, pose, and other modalities (Arunkumar and Ramasamy, 2021), these methods tend to ignore the potential complex relationships between modalities, resulting in poor performance of the models in cross-modal information fusion and emotion classification tasks. Therefore, leveraging the strengths of individual modalities while addressing the challenges of inter-modal alignment and information redundancy constitutes a critical challenge in advancing multimodal sentiment analysis.

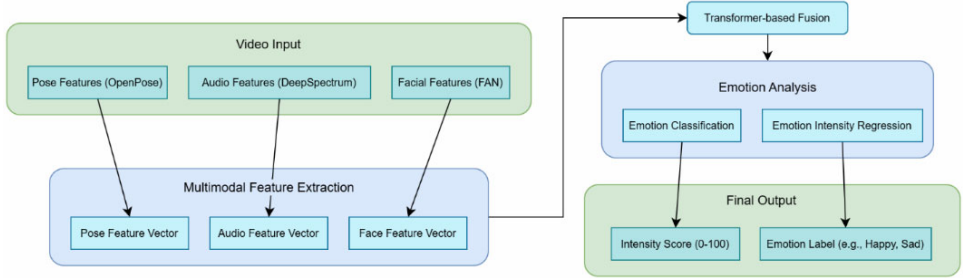
Based on the shortcomings and challenges of previous studies, this paper proposes a novel framework for multimodal sentiment analysis named multimodal emotion recognition with cross-attention transformers (MERCAT) (Sannidhan et al., 2023). Based on the multimodal feature extraction methods used in previous studies, this model combines multi-modal emotion feature extraction such as facial expression, audio, posture, etc. By incorporating the transformer architecture (Kodipalli et al., 2023) and a cross-modal self-attention mechanism into the fusion process, the model directly tackles the challenges of modal heterogeneity and inter-modal fusion.

The overall architectural design of the MERCAT model includes several important modules, each of which contributes significantly to the performance of sentiment analysis. First, the model extracts facial expression features through facial action network (FAN) to capture subtle changes in the face (Lee et al., 2024; Wenting et al., 2024), which provides rich emotional cues for sentiment recognition. Second, the DeepSpectrum model is used to extract audio features from videos to capture emotional information

(e.g., intonation, volume, speech rate, etc.) in speech. In addition, OpenPose was used to extract the pose features in the video to analyse the effect of body language on emotion (Cai et al., 2024; Fan et al., 2024). All extracted features are fed into the transformer network for fusion, which is able to capture the interrelationships between modalities through a stacked multi-head self-attention mechanism and generate a unified representation of emotion. Finally, the fully connected network (F-C) is responsible for categorising the fused features for emotion classification and intensity prediction (Li and Song, 2024), outputting the emotion categories (e.g., anger, joy, sadness, etc.) as well as the intensity values for each emotion.

Figure 1 shows the overall structure of this model. The figure shows the whole process of data from input to output, including the steps of feature extraction, information fusion and emotion prediction of data. Through this figure, we can clearly see the role of each module and the interaction between them.

Figure 1 The overall structure diagram of the model (see online version for colours)



In the process of network construction, the data is first input into the model through the preprocessing step, and after processing by the feature extraction module (facial, audio and pose feature extraction), it enters the transformer network. In transformer, the features of each modality are first weighted to calculate the correlation of emotions. Next, information is passed between modalities using a multi-head self-attention mechanism that fuses the features of each modality to generate a unified emotion representation. This sentiment representation is then passed to the fully-connected layer where sentiment classification or sentiment intensity prediction is performed. Through the joint training strategy, the model is able to accurately predict the intensity of sentiment while performing the sentiment classification task, thus making the sentiment analysis more comprehensive and detailed.

The model's principal advantage lies in its robust cross-modal fusion capability. This strength, enabled by the integration of the transformer architecture and a cross-modal self-attention mechanism, allows it to effectively capture inter-modal correlations and produce a more precise sentiment representation via weighted fusion. In addition, this model is able to simultaneously perform sentiment classification and intensity prediction with high accuracy and strong generalisation ability. The expected effect is that in multimodal sentiment recognition tasks, the present model can significantly improve the accuracy and robustness of sentiment analysis, especially in the prediction of sentiment intensity in complex environments, which shows superior performance.

Through this design, the present model can effectively deal with the problems of modal heterogeneity, information redundancy and sentiment intensity prediction faced by

traditional sentiment analysis models, and thus achieves more excellent performance in multimodal sentiment analysis tasks.

2.2 Multimodal feature integration

The multimodal feature fusion module in the MERCAT model is a key part of this study, which combines sentiment features from different modalities (e.g., facial expression, audio, and pose) to improve the accuracy and robustness of sentiment analysis (Pang et al., 2024; Goncalves and Busso, 2023). The basic principle of multimodal feature fusion lies in the fact that the information provided by different modalities complements each other to capture the expression of emotion more comprehensively (Antonio et al., 2024; Cao et al., 2024). Facial expressions reflect immediate expressions of emotion, audio reveals changes in emotional tone, and posture conveys additional emotional cues through body language (Lyu et al., 2024). These modalities have different feature spaces between them, but together they depict the complete emotional state. Therefore, how to efficiently fuse the information of these modalities becomes an important challenge in multimodal sentiment analysis (Wang et al., 2024; Zhou et al., 2024).

To achieve this goal, this paper proposes a cross-modal self-attention mechanism based on the transformer architecture, which aims to fuse the features between modalities and compute the relationship between them through the self-attention mechanism. Specifically, suppose we have multiple modal feature representations from facial expression, audio, and pose x_v , x_a and x_p , respectively, which are first extracted from their respective features by a pre-trained deep learning model. Next, these modalities are weighted and fused through the transformer model. The model's central component leverages self-attention to compute modality-specific weights, thereby generating a unified representation from diverse emotional signals.

In multimodal feature fusion, we first measure the importance of each modality in the expression of emotion by calculating the attentional weight α . The key to the self-attention mechanism is to compute the similarity between each feature and adjust its weight accordingly. This enables the model to dynamically assign varying contribution levels to each modality for sentiment analysis.

First, we calculate the attention scores α_v , α_a and α_p for each modality:

$$\alpha_m = \frac{\exp(f_m)}{\sum_{i \in M} \exp(f_i)} \quad (1)$$

where f_m is the similarity score of the features (e.g., facial, audio, or postural features) corresponding to the modality m and M denotes the set of all modalities. With this formula, we can obtain the importance weight α_m for each modality to weight the features.

Next, we fused the weighted features. By fusing the weighted features, we obtain a unified sentiment representation:

$$x_{fused} = \sum_{m \in M} \alpha_m x_m \quad (2)$$

where x_m denotes the feature representation of modality m and α_m is the weight corresponding to modality m . This weighted summation process ensures that the emotion

features of different modalities are effectively fused according to their importance, resulting in a more accurate emotion representation.

Finally, the fused features are passed through the fully connected layer for classification or regression to accomplish the sentiment analysis task. We use the classification or regression loss function to train the model and compute the final sentiment category and sentiment intensity:

$$L = \lambda L_{class} + (1 - \lambda) L_{reg} \quad (3)$$

where L_{class} is the loss function for the emotion categorisation task, L_{reg} is the loss function for the regression of the emotion intensity prediction task, and λ is a weighting coefficient for balancing the loss of the categorisation and regression tasks.

Through these core formulas, the model achieves effective multimodal weighting and fusion, thereby delivering accurate sentiment analysis via combined classification and regression.

In this model, the use of this cross-modal self-attention mechanism is able to effectively handle the heterogeneity among facial expressions, audio, and pose, and dynamically adjust the contribution of each modality during sentiment analysis. In this way, the model is able to automatically optimise the sentiment representation of each modality according to its ability to express emotion in different contexts, thus improving the accuracy and robustness of sentiment analysis.

2.3 *Emotion intensity prediction*

The sentiment intensity prediction task requires the model to be able to not only identify categories of sentiment, but also quantify the intensity of the sentiment (Ping and Yue, 2024). This is different from the traditional emotion classification task (Fenglin et al., 2023; Altun and Cordeiro, 2020), which only outputs the categories of emotions, whereas emotion intensity estimation requires more precise numerical outputs that involve fine-grained analysis of emotions (Wang et al., 2021; Choi and Boo, 2020). The core goal of sentiment intensity prediction is to regressively model the intensity of a sentiment through a multimodal fusion of sentiment representations, allowing the model to assign an intensity value to each sentiment category. An emotion intensity value is usually a continuous value indicating the intensity of the emotion, e.g., the intensity of an emotion can range from 0 to 100.

In the emotion intensity prediction task, this study used a regression model to handle the prediction of emotion intensity. The regression model differs from the classification model in that the output of the regression task is a continuous value instead of discrete category labels (Hu et al., 2022). Therefore, during model training, our goal is to minimise the error between the predicted value and the actual labelled value rather than performing category label matching.

First, we need to obtain a unified sentiment representation x_{fused} from the multimodal fusion module. This representation already contains sentiment information from multiple modalities such as facial expression, audio, and pose, and is weighted and fused by a cross-modal self-attention mechanism to ensure that the importance of each modality has been fully considered in the sentiment analysis. Next, this fused sentiment representation is passed to the regression network for sentiment intensity prediction.

In the regression network, the fused features x_{fused} are first mapped by inputting them into the fully-connected layer to compute the predicted values of sentiment intensity $\hat{y}$. The output we wish to obtain is the intensity value of the sentiment. For the regression task, the mean square error (MSE) serves as the loss function, with the objective of reducing the divergence between the prediction and the true label through parameter optimisation.

Specifically, assuming that the true sentiment intensity value is y , the loss function L_{reg} of the regression model can be expressed as:

$$L_{reg} = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2 \quad (4)$$

where N is the total number of samples, $\hat{y}_i$ is the value of emotional intensity predicted by the model, and y_i is the value of true emotional intensity. By minimising the mean square error, the model is able to learn how to map the multimodal sentiment representation to the exact sentiment intensity value.

In order to optimise the balance between the sentiment classification task and the sentiment intensity prediction task, this study combines the loss function L_{class} for sentiment classification with the regression loss function L_{reg} for co-training using a weighted loss function. The weighting coefficient λ controls the relative importance of the classification and regression tasks, and the final loss function L can be expressed as:

$$L = \lambda L_{class} + (1 - \lambda) L_{reg} \quad (5)$$

where the choice of λ depends on the dataset and task-specific requirements. Joint training facilitates a harmonious integration of the sentiment classification and intensity prediction tasks, leading to improved analytical accuracy and robustness.

In the process of sentiment intensity prediction, in addition to utilising the fused sentiment representation, the model also needs to consider the intensity distribution of different sentiment categories. In practical applications, the intensity ranges and representations of different emotions may vary. For example, the intensity changes of anger and joy may be more significant, while the intensity changes of fear and sadness may be more subdued. Therefore, how to flexibly deal with the intensity differences of different emotion categories in regression networks remains an issue worthy of in-depth exploration. By combining multimodal features and self-attention mechanism, the present model can effectively handle these differences and improve the accuracy of emotion intensity prediction.

In summary, sentiment intensity prediction plays a crucial role in this model. Through the regression analysis of multimodal sentiment representation, the model can not only identify the categories of sentiment, but also predict the intensity of sentiment, providing more detailed and accurate information for sentiment analysis. The key to this process lies in fusing the information of different modalities and accurately estimating the sentiment intensity through an optimised regression network, thus improving the quality and depth of sentiment analysis.

3 Results and discussion

3.1 Introduction to the datasets

AffectNet (Gera and Balasubramanian, 2021) is one of the most widely used facial expression datasets, focusing on the facial expression recognition task. The dataset comprises approximately 450,000 facial images. These images encompass the seven basic emotion categories: anger, disgust, fear, happiness, sadness, surprise, and neutral. The images come from different races, genders, and age groups, and are highly diverse and representative. On each image, the dataset provides not only the labelling of the emotion category, but also includes a rating of the intensity of the emotion. Due to its rich sample size and diversity, AffectNet plays an important role in facilitating the research of facial emotion analysis, and provides good data support for the visual modality feature extraction and emotion classification tasks in this study.

The Aff-Wild2 dataset (Kollias and Zafeiriou, 2019), on the other hand, is a large dataset focusing on sentiment analysis in natural environments, containing thousands of long video clips with facial expression annotations. These videos are derived from natural scenes and contain labels for six basic emotions and neutral expressions. Unlike traditional static image datasets, Aff-Wild2's samples have a long temporal order, which makes it more suitable for multimodal sentiment analysis tasks, especially in video sentiment analysis, and better reflects the dynamic changes in emotions. This dataset is characterised by not only facial expression changes in the video, but also speech, background noise and other information, which provides a good basis for multimodal sentiment analysis. The application of the Aff-Wild2 dataset is mainly used in this study for feature extraction of audio, visual and postural modalities, as well as for temporal sequential modelling of sentiment changes.

These two datasets provide very comprehensive training and testing data for this study, covering all aspects of sentiment classification and sentiment intensity prediction tasks. By combining these two datasets, we are able to validate the performance of the model in different sentiment expressions, different data sources, and multimodal fusion tasks, thus ensuring that the proposed multimodal sentiment analysis model has strong robustness and generalisation capabilities. Overall, the diversity and high-quality labelling of these datasets provide the necessary experimental foundation for this study, helping us to deeply investigate and validate the effectiveness of multimodal sentiment analysis.

3.2 Experimental setup and evaluation metrics

To comprehensively evaluate the performance of the proposed multimodal sentiment analysis model, this section elaborates on the experimental setup and evaluation metrics. The model's effectiveness in both sentiment classification and intensity prediction is validated through comparative experiments, employing standard metrics to assess its accuracy and robustness.

3.2.1 Experimental setup

In this experiment, we use three multimodal sentiment analysis datasets: AffectNet and Aff-Wild2. For each dataset, we first perform data preprocessing to ensure data quality

and consistency. Specifically, we performed frame extraction on the video data and normalised the images so that each frame input to the model has the same resolution and color channel. For audio data, we used the DeepSpectrum method for feature extraction to obtain emotionally relevant features for each audio segment.

To evaluate the model's generalisation capability during training, we employed a cross-validation strategy. The dataset was partitioned into training, validation, and test sets following a standard split of 80%, 10%, and 10%, respectively. We employed the Adam optimiser to optimise the model parameters, with the learning rate set to 0.0001, and used the early stopping method to prevent overfitting. A batch size of 32 and training epochs of 50 were used for training to ensure that the model learns sufficiently and reaches the optimal state.

3.2.2 Experimental procedure

When conducting experiments, we first extract facial, audio and pose features by preprocessing the dataset and input them into the model for training. After each training cycle, the model is evaluated using the validation set and the model parameters are adjusted according to the evaluation results. The model's final performance is assessed on the test set, yielding results for both sentiment classification and intensity prediction. Each metric is computed individually to provide a comprehensive evaluation.

3.2.3 Evaluation metrics

To comprehensively evaluate the performance of this model in sentiment classification and sentiment intensity prediction tasks, we selected several standard evaluation metrics.

- 1 Accuracy: for sentiment classification tasks, accuracy is one of the most commonly used metrics to measure model classification performance. Accuracy represents the proportion of sentiment categories correctly predicted by the model to the total number of predictions. The calculation formula is:

$$L = \lambda L_{class} + (1 - \lambda) L_{reg} \quad (6)$$

- 2 F1-score: on imbalanced datasets, accuracy may not fully reflect model performance. Therefore, we also used the F1-score to comprehensively consider the model's precision and recall. The F1-score is calculated as follows:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

Precision refers to the proportion of samples predicted to belong to a particular sentiment category that actually belong to that category; recall refers to the proportion of samples that actually belong to a particular sentiment category that are correctly predicted to belong to that category. A higher F1-score indicates better overall model performance in sentiment classification tasks.

- 3 Mean squared error (MSE): for the sentiment intensity prediction task, we use MSE as the evaluation metric. MSE measures the difference between the predicted and true values. The calculation formula is:

$$MSE = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2 \quad (8)$$

where $\hat{y}_i$ is the model's predicted sentiment intensity value, y_i is the true sentiment intensity value, and N is the number of samples. A smaller MSE value indicates a better model performance in sentiment intensity prediction.

- 4 Pearson correlation coefficient (PCC): the Pearson correlation coefficient measures the linear correlation between the predicted and true sentiment intensity values. PCC values range from -1 to 1 , with values closer to 1 indicating a stronger correlation between the predicted and true results. The calculation formula is:

$$PCC = \frac{\sum_{i=1}^N (\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^N (\hat{y}_i - \bar{\hat{y}})^2 \sum_{i=1}^N (y_i - \bar{y})^2}} \quad (9)$$

where $\bar{\hat{y}}$ and $\bar{y}$ are the means of the predicted and actual values, respectively. PCC measures the linear relationship between sentiment intensity predictions. Values closer to 1 indicate greater consistency between the model's predictions and the actual sentiment intensity.

3.3 Experimental results

Our work rigorously evaluates the proposed multimodal sentiment analysis model, MERCAT. The evaluation employs three datasets (AffectNet, Aff-Wild2) to benchmark the model's capabilities in sentiment categorisation and intensity estimation against existing approaches. The experimental outcomes are detailed below, followed by a thorough analysis of each table.

3.3.1 Experimental results for the emotion categorisation task

A comparative evaluation of the MERCAT model and five benchmarks on the sentiment classification task is summarised in Table 1. Model performance is gauged using five evaluation metrics: accuracy, F1-score, precision, recall and AUC, offering a comprehensive view of their categorisation prowess.

Table 1 Performance comparison of emotion classification task models

<i>Model</i>	<i>Accuracy</i>	<i>F1-score</i>	<i>Precision</i>	<i>Recall</i>	<i>AUC</i>
MERCAT	91.6	0.89	0.88	0.9	0.92
FaceCNN (Seo and Chung, 2019)	87.2	0.83	0.81	0.85	0.86
AudioCNN (Krishna et al., 2024)	82.1	0.79	0.78	0.8	0.82
PoseCNN (Wang et al., 2023)	80.3	0.77	0.75	0.79	0.8
Early fusion CNN (Al-Jabbar et al., 2023)	85.6	0.82	0.8	0.83	0.84
Late fusion RNN (Le et al., 2021)	84.4	0.8	0.78	0.82	0.83

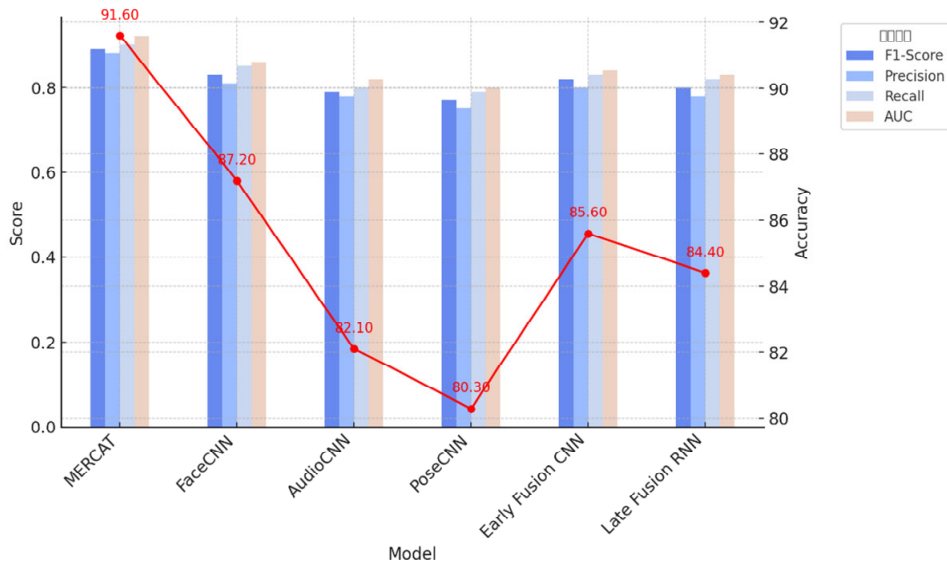
As can be seen in Table 1, the MERCAT model achieves the best results in all evaluation metrics, especially in accuracy (91.6%) and F1-score (0.89), which are significantly

better than the other models. Compared with FaceCNN (87.2% accuracy and 0.83 F1-score), MERCAT improves the accuracy by about 4.4% and F1-score by 6 percentage points, indicating that multimodal feature fusion can significantly improve the accuracy of emotion classification. For AudioCNN and PoseCNN, the improvement of MERCAT is even more obvious, which also indicates the limitation of unimodal models in processing complex emotions, especially in the sentiment analysis of multimodal data, where the combination of face, audio and pose can provide more comprehensive information.

In addition, the AUC value (0.92) of MERCAT is higher than all the benchmark models, indicating that the model has a better ability to differentiate between emotion categories. This also suggests that when dealing with the task of emotion categorisation in videos, MERCAT is able to more accurately differentiate between different emotion categories, especially the harder to differentiate emotions such as sadness and anger.

In order to visualise the performance of the MERCAT model in the sentiment classification task, we visualise the experimental results as a combination of bar charts and line graphs. Figure 1 shows the various evaluation metrics of different models in the sentiment classification task. Through this figure, we can see that the MERCAT is able to effectively fuse facial, audio, and pose features through the cross-modal self-attention mechanism and the transformer architecture, thus significantly improving the accuracy of emotion classification. Compared with other benchmark models, MERCAT shows significant advantages.

Figure 2 Comparison of model performance for emotion classification tasks (see online version for colours)



As can be seen in Figure 2, MERCAT is able to effectively fuse facial, audio, and pose features through the cross-modal self-attention mechanism and the transformer architecture, thus significantly improving the accuracy of emotion classification. Compared with other benchmark models, MERCAT shows significant advantages.

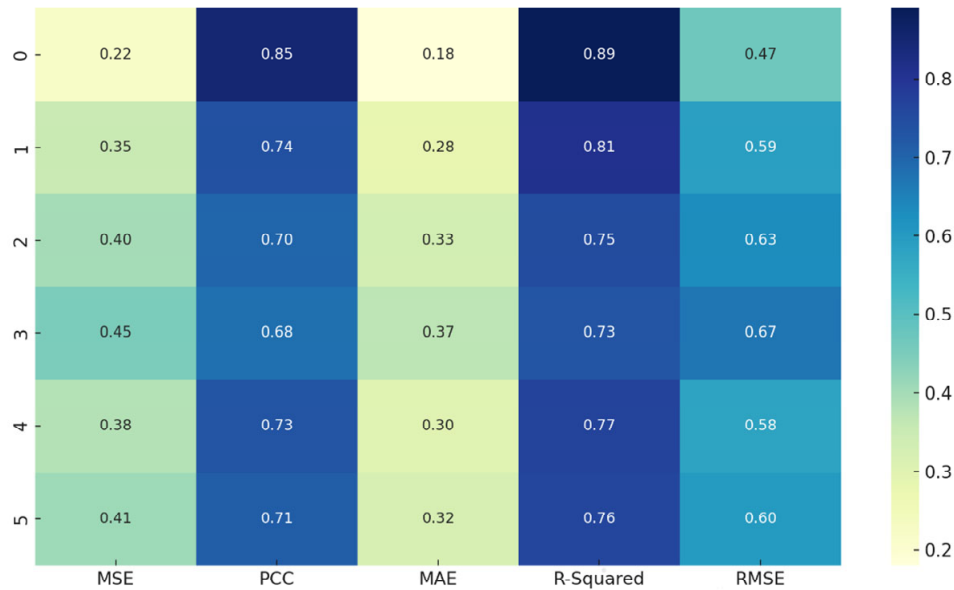
3.3.2 Experimental results for the emotional intensity prediction task

Table 2 demonstrates the performance of the MERCAT model in the emotion intensity prediction task and compares it with other benchmark models. The evaluation metrics include mean square error (MSE), Pearson correlation coefficient (PCC), mean absolute error (MAE), R-squared (coefficient of determination), and root mean square error (RMSE). These metrics provide a comprehensive measure of the model’s accuracy and relevance in the task of emotion intensity prediction.

Table 2 Comparison of performance of emotion intensity prediction task models

Model	MSE	PCC	MAE	R-squared	RMSE
MERCAT	0.22	0.85	0.18	0.89	0.47
FaceCNN	0.35	0.74	0.28	0.81	0.59
AudioCNN	0.4	0.7	0.33	0.75	0.63
PoseCNN	0.45	0.68	0.37	0.73	0.67
Early fusion CNN	0.38	0.73	0.3	0.77	0.58
Late fusion RNN	0.41	0.71	0.32	0.76	0.6

Figure 3 Comparison of model performance in emotion intensity prediction tasks (see online version for colours)



As can be seen from Table 2, the MERCAT model also shows significant advantages in the emotion intensity prediction task. The MSE of MERCAT (0.22) is much lower than that of the other benchmark models, especially FaceCNN (MSE 0.35) and AudioCNN (MSE 0.40), which suggests that MERCAT has a smaller error in emotion intensity prediction and can provide more accurate estimation of emotion intensity. Also, MERCAT’s PCC (0.85) is higher than the other models, showing its strong predictive

correlation, which implies a closer relationship between the model-predicted affective intensity and the actual intensity values.

For the MAE (0.18) and RMSE (0.47) metrics, MERCAT also outperforms the other benchmark models, especially in terms of prediction accuracy, indicating that the model is able to more accurately predict the intensity of the emotion with less error. Compared to PoseCNN and early fusion CNN, MERCAT performs more prominently, further demonstrating the effectiveness of cross-modal feature fusion and self-attention mechanisms in emotion intensity prediction.

In the emotion intensity prediction task, we used heat maps to show the performance of different models on each evaluation metric. Figure 3 demonstrates the comparison between MERCAT and other benchmark models on each assessment metric. It can be seen that MERCAT has an absolute advantage in the emotion intensity prediction task, especially after fusion using multimodal features, the model is able to capture the intensity change of emotion more accurately. By combining facial, audio, and postural features, MERCAT is able to effectively reduce errors and improve the accuracy of emotion intensity prediction.

From the analysis of the results of the above two experiments, it can be seen that the MERCAT model shows significant advantages in both sentiment classification and sentiment intensity prediction tasks. Multimodal information fusion plays a very important role in the task of sentiment analysis, especially in complex emotional expressions and dynamically changing emotional situations, fusing features from different modalities (e.g., facial expression, audio, and pose) can provide a more comprehensive and accurate representation of sentiment. Through the transformer architecture and cross-modal self-attention mechanism, MERCAT can effectively handle the heterogeneity of multimodal data and enhance the performance of sentiment analysis by learning the correlations between modalities.

MERCAT exhibits across-the-board improvements over unimodal models in all metrics. The integration of multimodal data provides a more comprehensive information base, which is crucial for accurately interpreting complex sentiments and temporal dynamics, resulting in a more detailed and reliable analysis.

The experimental findings collectively demonstrate MERCAT's proficiency and robust generalisation in multimodal sentiment analysis. Promising avenues for future work include improving the model's efficacy in scenarios involving extreme emotions and longitudinal emotion shifts, as well as boosting its computational performance.

4 Conclusions

In this study, we propose a multimodal sentiment feature extraction and information fusion method for video content, aiming to improve the accuracy and robustness of sentiment analysis. Sentiment analysis, as a key task in the field of affective computing, faces the challenge of multimodal data fusion, especially in video content, where information such as facial expression, audio, and pose play an important role in expressing emotions. Traditional unimodal sentiment analysis methods often neglect the interaction information between different modalities, resulting in limited accuracy of sentiment analysis. To address this problem, our proposed MERCAT model effectively captures inter-modal interrelationships by fusing multimodal features in videos,

combined with the transformer architecture and cross-modal self-attention mechanism, which improves the performance of sentiment classification and intensity prediction.

The main contribution of this study is to propose an innovative multimodal sentiment analysis model that is able to enhance the performance of sentiment analysis through self-attention mechanism and cross-modal information fusion. Through experimental validation, the MERCAT model shows significant advantages in both sentiment classification and intensity prediction tasks, providing a new solution idea in the field of multimodal sentiment analysis. However, this study also has certain shortcomings. First, the model may still face the problem of decreasing sentiment recognition accuracy when dealing with complex sentiment scenes, especially for the expression of extreme emotions, the model may not adequately capture their nuances. Second, the computational efficiency and training time of the MERCAT model when processing large-scale video data still need to be further optimised. Future research could focus on how to improve the training speed and computational efficiency of the model, especially for applications on large-scale datasets.

Future research efforts will focus on the following directions. First, we plan to introduce more advanced temporal modelling methods to further improve the performance of the model when dealing with dynamic emotion changes, especially when facing videos with long time periods. Second, we hope to improve the model's architecture and algorithms to enhance its ability to recognise extreme emotions and fine-grained emotion changes. In addition, future work will focus on how to reduce the computational cost and training time of the model, and improve the efficiency of the model in practical applications by means of model compression and hardware acceleration. With these improvements, the MERCAT model is expected to play an important role in a wider range of sentiment analysis applications.

Declarations

The author declares that he has no conflicts of interest.

References

- Abinaya, P. and Kumar, S.J. (2024) 'Assured and provable data expuncturing in cloud using ciphertext policy-attribute based encryption (CP-ABE)', *Cybernetics and Systems*, Vol. 55, No. 4, pp.786–803.
- Al-Jabbar, M., Alshahrani, M., Senan, E.M. and Ahmed, I.A. (2023) 'Analyzing histological images using hybrid techniques for early detection of multi-class breast cancer based on fusion features of CNN and handcrafted', *Diagnostics*, Vol. 13, No. 10, p.1753.
- Altun, E. and Cordeiro, G.M. (2020) 'The unit-improved second-degree Lindley distribution: inference and regression modeling', *Computational Statistics*, Vol. 35, pp.259–279.
- Antonio, G., Jose, J., George, S.N. and Raja, K. (2024) 'CroMA: cross-modal attention for visual question answering in robotic surgery', *International Conference on Pattern Recognition*, Springer, pp.459–471.
- Arunkumar, P. and Ramasamy, L.K. (2021) 'Time-series forecasting and analysis of COVID-19 outbreak in highly populated countries: a data-driven approach', *International Journal of E-Health and Medical Communications (IJEHMC)*, Vol. 13, No. 2, pp.1–17.

- Cai, J., Sharkawi, I. and Taasim, S.I. (2024) 'The role of deep learning in supply chain collaboration and cooperation', *Journal of Organizational and End User Computing (JOEUC)*, Vol. 36, No. 1, pp.1–23.
- Cao, J., Zou, Q., Ning, X., Wang, Z. and Wang, G. (2025) 'FIPNet: self-supervised low-light image enhancement combining feature and illumination priors', *Neurocomputing*, Vol. 623, p.129426.
- Cao, S., An, G., Cen, Y., Yang, Z. and Lin, W. (2024) 'CAST: cross-modal retrieval and visual conditioning for image captioning', *Pattern Recognition*, Vol. 153, p.110555.
- Choi, Y. and Boo, Y. (2020) 'Comparing logistic regression models with alternative machine learning methods to predict the risk of drug intoxication mortality', *International Journal of Environmental Research and Public Health*, Vol. 17, No. 3, p.897.
- DeMatteo, C., Jakubowski, J., Stazyk, K., Randall, S., Perrotta, S. and Zhang, R. (2024) 'The headaches of developing a concussion app for youth: balancing clinical goals and technology', *International Journal of E-Health and Medical Communications (IJEHMC)*, Vol. 15, No. 1, pp.1–20.
- Fan, L., Wang, C. and Lu, Z. (2024) 'Application of AdaBound-optimized XGBoost-LSTM model for consumer credit assessment in banking industries', *Journal of Organizational and End User Computing (JOEUC)*, Vol. 36, No. 1, pp.1–24.
- Fenglin, W. et al. (2023) 'Exploratory regression modeling for flood susceptibility mapping in the GIS environment', *Scientific Reports*, Vol. 13, No. 1, p.247.
- Gera, D. and Balasubramanian, S. (2021) 'Landmark guidance independent spatio-channel attention and complementary context information based facial expression recognition', *Pattern Recognition Letters*, Vol. 145, pp.58–66.
- Goncalves, L. and Busso, C. (2023) 'Learning cross-modal audiovisual representations with ladder networks for emotion recognition', *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp.1–5.
- Hu, Y., Satten, G.A. and Hu, Y-J. (2022) 'LOCOM: a logistic regression model for testing differential abundance in compositional microbiome data with false discovery rate control', *Proceedings of the National Academy of Sciences*, Vol. 119, No. 30, p.e2122788119.
- Huang, J. et al., (2025) 'Design and verification of a wearable micro-capacitance test system for POC biosensing', *IEEE Transactions on Instrumentation and Measurement*, Vol. 74, pp.1–11.
- Huang, J., Yu, X., An, D., Ning, X., Liu, J. and Tiwari, P. (2025) 'Uniformity and deformation: a benchmark for multi-fish real-time tracking in the farming', *Expert Systems with Applications*, Vol. 264, p.125653, <https://doi.org/10.1016/j.eswa.2024.125653>.
- Jawahar, M., Anbarasi, L.J., Jayachandran, P., Ramachandran, M. and Al-Turjman, F. (2021) 'Utilization of transfer learning model in detecting COVID-19 cases from chest x-ray images', *International Journal of E-Health and Medical Communications (IJEHMC)*, Vol. 13, No. 2, pp.1–11.
- Kodipalli, A., Fernandes, S.L., Dasar, S.K. and Ismail, T. (2023) 'Computational framework of inverted fuzzy C-means and quantum convolutional neural network towards accurate detection of ovarian tumors', *International Journal of E-Health and Medical Communications (IJEHMC)*, Vol. 14, No. 1, pp.1–16.
- Kollias, D. and Zafeiriou, S. (2019) *Expression, Affect, Action Unit Recognition: Aff-wild2, Multi-Task Learning and Arcface*, *arXiv preprint arXiv:1910.04855*.
- Krishna, N.B., Reddy, R.S.V., Likhith, M. and Priya, N.L. (2024) 'Tackling depression detection with deep learning: a hybrid model', *Driving Smart Medical Diagnosis through AI-Powered Technologies and Applications*, IGI Global, pp.102–117, Chikaballapur, India.
- Le, D., Keren, G., Chan, J., Mahadeokar, J., Fuegen, C. and Seltzer, M.L. (2021) 'Deep shallow fusion for RNN-T personalization', *2021 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, pp.251–257.

- Lee, P-H., Sun, Y-K., Ke, Y-P. and Lee, P-J. (2024) 'Hotel rating prediction system based on time factors: using reviews and sentiment analysis', *Journal of Organizational and End User Computing (JOEUC)*, Vol. 36, No. 1, pp.1–29.
- Li, J., Wang, Y., Ning, X., He, W. and Cai, W. (2025) 'FefDM-Transformer: dual-channel multi-stage transformer-based encoding and fusion mode for infrared-visible images', *Expert Systems with Applications*, Vol. 277, p.127229.
- Li, M. and Guenier, A.W. (2024) 'ChatGPT and health communication: a systematic literature review', *International Journal of E-Health and Medical Communications (IJEHMC)*, Vol. 15, No. 1, pp.1–26.
- Li, T. and Song, J. (2024) 'Deep learning-powered financial product recommendation system in banks: integration of transformer and transfer learning', *Journal of Organizational and End User Computing (JOEUC)*, Vol. 36, No. 1, pp.1–29.
- Lyu, T. et al. (2024) *Optimized CNNs for Rapid 3D Point Cloud Object Recognition*, arXiv preprint arXiv:2412.02855.
- Meng, Z. and Lin, M. (2023) 'The driving factors analysis of live streamers' characteristics and perceived value for consumer repurchase intention on live streaming platforms', *Journal of Organizational and End User Computing (JOEUC)*, Vol. 35, No. 1, pp.1–24.
- Ning, X. et al. (2025) 'ABM: an automatic body measurement framework via body deformation and topology-aware B-spline approximation', *Pattern Recognition*, Vol. 277, p.112060.
- Pang, H. et al. (2024) *Electronic Health Records-Based Data-Driven Diabetes Knowledge Unveiling and Risk Prognosis*, arXiv preprint arXiv:2412.03961.
- Ping, R.B. and Yue, W.Z. (2024) 'Strategic focus, tasks, and pathways for promoting china's modernization through new productive forces', *Journal of Xi'an University of Finance and Economics*, Vol. 1, No. 12, pp.3–11.
- Poloju, N. and Rajaram, A. (2024) 'Transformation with YOLO tiny network architecture for multimodal fusion in lung disease classification', *Cybernetics and Systems*, pp.1–22.
- Qi, H. et al. (2022) 'Capacitive aptasensor coupled with microfluidic enrichment for real-time detection of trace SARS-CoV-2 nucleocapsid protein', *Analytical chemistry*, Vol. 94, No. 6, pp.2812–2819.
- Sannidhan, M., Martis, J.E., Nayak, R.S., Aithal, S.K. and Sudeepa, K. (2023) 'Detection of antibiotic constituent in *Aspergillus flavus* using quantum convolutional neural network', *International Journal of E-Health and Medical Communications (IJEHMC)*, Vol. 14, No. 1, pp.1–26.
- Seo, J. and Chung, I-J. (2019) 'Face liveness detection using thermal face-CNN with external knowledge', *Symmetry*, Vol. 11, No. 3, p.360.
- Wang, J., Li, F. and He, L. (2025a) 'A unified framework for adversarial patch attacks against visual 3D object detection in autonomous driving', *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 35, No. 5, pp.4949–4962.
- Wang, J., Li, F., Lv, S., He, L. and Shen, C. (2025b) 'Physically realizable adversarial creating attack against vision-based BEV space 3D object detection', *IEEE Transactions on Image Processing*, Vol. 34, pp.538–551.
- Wang, S., Jiang, R., Wang, Z. and Zhou, Y. (2024) *Deep Learning-Based Anomaly Detection and Log Analysis for Computer Networks*, arXiv preprint arXiv:2407.05639.
- Wang, S., Ning, Y. and Shi, H. (2021) 'A new uncertain linear regression model based on equation deformation', *Soft Computing*, Vol. 25, No. 20, pp.12817–12824.
- Wang, S., Zhang, Y. and Zhuang, Y. (2023) *Closing the Loop: Single-Image to Camera Pose with Inverse NeRF and PoseCNN*.
- Wenting, L., Hussain, W.M.H.W., Xinlin, J., Na, M. and Alam, S.S. (2024) 'Analyzing the impact on talent acquisition and performance management: HR and data analysis', *Journal of Organizational and End User Computing (JOEUC)*, Vol. 36, No. 1, pp.1–30.

- Xing, B. et al. (2019) 'Exploiting EEG signals and audiovisual feature fusion for video emotion recognition', *IEEE Access*, Vol. 7, pp.59844–59861.
- Xing, X., Wang, B., Ning, X., Wang, G. and Tiwari, P. (2025) 'Short-term OD flow prediction for urban rail transit control: a multi-graph spatiotemporal fusion approach', *Information Fusion*, Vol. 118, No. 06, p.102950, <https://doi.org/10.1016/j.inffus.2025.102950>.
- Zeng, L. (2023) 'The influence of knowledge worker salary satisfaction on employee job performance', *Journal of Organizational and End User Computing (JOEUC)*, Vol. 35, No. 1, pp.1–17.
- Zhou, Y. et al. (2024) 'Optimization of automated garbage recognition model based on ResNet-50 and weakly supervised CNN for sustainable urban development', *Alexandria Engineering Journal*, Vol. 108, pp.415–427.