

Article

MMER-LMF: Multi-Modal Emotion Recognition in Lightweight Modality Fusion

Eun-Hee Kim ¹, Myung-Jin Lim ²  and Ju-Hyun Shin ^{2,*} ¹ Department of Computer Science, Chosun University, Gwangju 61452, Republic of Korea; eunflash00@chosun.ac.kr² Department of Future Convergence, Chosun University, Gwangju 61452, Republic of Korea; myungjin@chosun.ac.kr

* Correspondence: jhshinkr@chosun.ac.kr

Abstract: Recently, multimodal approaches that combine various modalities have been attracting attention to recognizing emotions more accurately. Although multimodal fusion delivers strong performance, it is computationally intensive and difficult to handle in real time. In addition, there is a fundamental lack of large-scale emotional datasets for learning. In particular, Korean emotional datasets have fewer resources available than English-speaking datasets, thereby limiting the generalization capability of emotion recognition models. In this study, we propose a more lightweight modality fusion method, MMER-LMF, to overcome the lack of Korean emotional datasets and improve emotional recognition performance while reducing model training complexity. To this end, we suggest three algorithms that fuse emotion scores based on the reliability of each model, including text emotion scores extracted using a pre-trained large-scale language model and video emotion scores extracted based on a 3D CNN model. Each algorithm showed similar classification performance except for slight differences in disgust emotion performance with confidence-based weight adjustment, correlation coefficient utilization, and the Dempster–Shafer Theory-based combination method. The accuracy was 80% and the recall was 79%, which is higher than 58% using text modality and 72% using video modality. This is a superior result in terms of learning complexity and performance compared to previous studies using Korean datasets.



Academic Editor: Yue Wu

Received: 15 April 2025

Revised: 9 May 2025

Accepted: 23 May 2025

Published: 24 May 2025

Citation: Kim, E.-H.; Lim, M.-J.; Shin, J.-H. MMER-LMF: Multi-Modal Emotion Recognition in Lightweight Modality Fusion. *Electronics* **2025**, *14*, 2139. <https://doi.org/10.3390/electronics14112139>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: multi-modal emotion recognition; decision-level fusion; text–video

1. Introduction

As human–robot interaction (HRI) becomes increasingly important due to advancements in artificial intelligence (AI) and robotics, emotion recognition technology plays a critical role in enabling natural and seamless communication between humans and machines across various fields such as entertainment, healthcare, and education [1]. For instance, if virtual assistants or social robots can accurately identify users’ emotional states, they can provide more personalized and context-aware responses. This has led to a growing body of research aimed at developing accurate emotion recognition systems [2].

Traditionally, emotion recognition has relied on single modalities such as voice, facial expressions, or physiological signals. While these approaches have been widely studied, they often fail to fully capture the complexity of human emotions. Human emotions are inherently complex due to their ambiguous, overlapping, temporally dynamic, and culturally influenced nature. For instance, emotions like fear and surprise can share similar facial expressions, while anger and disgust may be indistinguishable from vocal tones alone.

Furthermore, emotional expressions are often context-dependent, requiring integration of both verbal and nonverbal cues for accurate interpretation. As a result, single-modal systems—whether based on facial expressions, speech, or text—struggle to resolve these ambiguities and often produce inconsistent or incomplete emotional assessments. This has sparked increasing interest in multimodal methods that integrate text, video, and audio to enhance recognition accuracy [3–5]. However, while multimodal fusion offers high accuracy, it is challenging to apply in real-time systems or on lightweight devices due to its high computational and memory demands. This difficulty stems from the heterogeneity of multimodal data, which makes it hard to integrate different types of information effectively. Since data fusion plays a key role in multimodal emotion recognition, the overall model performance largely depends on how well these modalities are combined [6]. Feature-level fusion, also known as early fusion, combines data from different modalities at the input stage and trains a single model, allowing it to capture inter-modal correlations from the beginning [7]. However, it often suffers from issues such as high dimensionality and difficulties in time synchronization.

In contrast, decision-level fusion—also referred to as late fusion—processes each modality separately and integrates the results at the final stage. This approach provides flexibility and allows independent learning for each modality, but it can be challenging to design effective fusion rules and may suffer from limited performance due to overly simplistic combination strategies.

Recent studies have explored and compared various fusion techniques to identify the most effective strategies for multimodal emotion recognition. With the rapid advancement of deep learning, its application to multimodal data fusion has become increasingly widespread [8,9]. For instance, Cimtay et al. [10] demonstrated that hybrid fusion methods, which integrate the strengths of both feature-level and decision-level fusion, can significantly improve classification accuracy. Sahu and Vechtomova [11] proposed an adaptive fusion technique that enables neural networks to dynamically determine the optimal way to combine multimodal inputs, outperforming traditional fusion approaches. In a related domain, Li et al. [12] introduced a sensor-adaptive multimodal fusion method for 3D object detection, effectively combining LiDAR and camera data through a deformable attention mechanism to produce a unified representation across modalities.

Despite recent advances, securing large-scale emotional datasets remains a fundamental challenge in affective computing. Most studies rely on relatively small corpora, which limits both the learning capacity and generalization ability of emotion recognition models. This problem is particularly acute in Korean-language research, where high-quality multimodal emotion datasets—especially video-based—are significantly fewer than in English. To address this underrepresentation, we focus on a Korean-language multimodal video dataset in this study. Importantly, our aim is not to develop a language-specific system but to propose a modular fusion architecture that generalizes across languages. The proposed framework utilizes transferable features such as text embeddings and visual emotion dynamics, enabling potential deployment in multilingual contexts.

In low-resource scenarios, feature-level multimodal fusion is especially challenging due to the high dimensionality and synchronization demands of heterogeneous inputs. To overcome this, we propose a decision-level fusion framework that integrates modality-specific emotion scores from text and video. While recent advancements in pre-trained language models—particularly GPT-based generative models—have improved contextual and emotional understanding in text, their performance still suffers from domain shifts, ambiguous expressions, and the absence of nonverbal cues. Emotional meaning in short utterances often remains unclear, and text cannot capture critical signals such as tone, facial expressions, or gestures.

Therefore, for more robust and context-aware emotion recognition, a multimodal approach that integrates both verbal and nonverbal information is essential. By combining the strengths of pretrained language models with lightweight visual emotion classification and efficient decision-level fusion, our framework aims to balance accuracy, adaptability, and computational efficiency. Therefore, this study aims to enhance emotion recognition by addressing the following key aspects:

- Utilization of pre-trained language models: To overcome the limited availability of Korean emotion datasets, we explore the use of large-scale pre-trained models for text-based emotion analysis.
- Multimodal fusion: By integrating text and video data, we aim to mitigate the limitations of single-modality approaches and improve recognition accuracy.
- Emotion score-based decision fusion: We propose various fusion strategies that leverage the emotion scores from each modality, accounting for differences in classification performance.

The structure of this paper is as follows: Section 2 reviews related work. Section 3 details the proposed methodology. Section 4 presents the experimental setup, results, and comparative analysis. Section 5 concludes the paper with a summary and future research directions.

2. Related Work

This section reviews previous studies that have employed feature-level, decision-level, and hybrid fusion techniques for multimodal emotion recognition, particularly in video-based contexts.

Viegas et al. [13] proposed a two-stage emotion recognition model that combines frame-level and video-level features. It applied feature-level fusion by combining Facial Action Units (FAUs), voice features (MFCC), and skeleton motion data, and utilized a method of classifying emotions into stage 1 (positive/neutral/negative) and stage 2 (detailed emotions). The experimental results showed that video-level features outperform frame-level features, suggesting that a fusion method that reflects temporal emotional changes can improve emotion recognition performance.

Kumar et al. [14] proposed an audio-video multimodal emotion recognition model using deep learning and applied a feature-level fusion method that combined voice feature extraction using ResNet20 and image feature analysis using 3D CNN. In particular, OpenSMILE was used to extract voice features (MFCC, pitch, energy, etc.), facial emotions were analyzed from the image, and the Weighted Concatenation Fusion technique was applied to generate the optimal emotion expression vector. However, the feature-level fusion method may be difficult to apply in real applications due to high-dimensional feature problems and time synchronization issues, and, in particular, generalization performance may be degraded when the dataset is insufficient. In Lee Dae-jong's study [15], a fusion method was applied that combined facial features extracted from still images with temporal emotional changes analyzed from videos. In particular, we effectively analyzed emotional changes that are difficult to detect in still images by utilizing a dynamic modeling technique based on the HMM (Hidden Markov Model), thereby enabling more precise emotion recognition. However, if the emotional features of still images and videos are not consistent, the problem of inter-modality imbalance may occur, and especially if the dataset size is insufficient, the model performance may deteriorate.

Jo and Jung's study [16] proposed a multimodal emotion recognition system that combines facial images and multidimensional emotion-based text information. This method applies a fusion method that combines a CNN-based facial emotion recognition model and a BERT-based text emotion recognition model to quantitatively analyze emotions in a multidimensional space (VAD: Valence–Arousal–Dominance). In particular, by combin-

ing feature-level fusion and decision-level fusion, they optimized the integration of the emotional information of facial images and texts, effectively improving the accuracy of emotional recognition compared to single-modality approaches. However, text emotion analysis has limitations. It is difficult for learned models to consistently perform in real-world situations due to differences in speaking styles across domains, and it is difficult to recognize emotions when contextual information is lacking. S. Moon's study [17] proposed a multimodal method for the emotion classification of Korean video data. They designed deep learning models specialized for video and voice. The video model extracts features using a 3D convolutional neural network, and the voice model uses MFCC technology to create spectrogram images that can visualize and express voice data. The feature vectors from the two feature extraction methods were merged, and the emotion with the highest probability was predicted using the Softmax function. However, since the Korean emotional dataset is limited, the generalization performance of the trained model is likely to be low, and additional fusion strategies are needed to address the problem of emotional information mismatch between voice and video. Ortega et al.'s study [18] proposed a multimodal emotion recognition model using deep neural networks. Unlike previous studies that applied the Early Fusion (feature combination) or Late Fusion (result combination) method, we introduced a new approach that learns audio, video, and text data individually and then learns the optimal emotional expression using the Intermediate Fusion method. The proposed model applies a method in which each modality is learned in independent layers and then combined in shared layers to perform emotion classification. However, although the Intermediate Fusion method recorded high performance, it has limitations in that its performance may deteriorate when there is not much training data, and it is difficult to apply it to real-time emotion analysis.

In the study by Yi et al. [19], the HyFusER (Hybrid Multimodal Transformer for Emotion Recognition Using Dual Cross-Modal Attention) model was proposed to improve emotion recognition performance. This model extracts text features using the KoELECTRA pre-trained model, extracts speech features using HuBERT, then learns features of both modalities using the Transformer encoder, and then effectively combines information between the two modalities using the Dual Cross Modal Attention technique. In particular, information exchange was maximized by applying Intermediate Layer Fusion and Last Fusion, and the final emotion prediction was performed using the average ensemble method. However, because it uses various pre-learning models and complex fusion techniques, it has the limitation that the model structure is complex and the amount of computational processing is large, and the performance is not high compared to that.

Recent efforts in lightweight emotion recognition have also adopted entropy-based feature fusion to minimize computational overhead. For example, Li et al. [20] proposed a resource-efficient EEG emotion recognition method that converts single-channel signals into a Brain Rhythm Entropy Matrix (BREM) using multi-entropy features. This matrix is then classified using simple similarity measures, achieving over 80% accuracy on the DEAP dataset while maintaining interpretability and reducing model complexity. Such entropy-driven approaches provide a promising direction for real-time applications in constrained environments, aligning with our goal of lightweight multimodal systems.

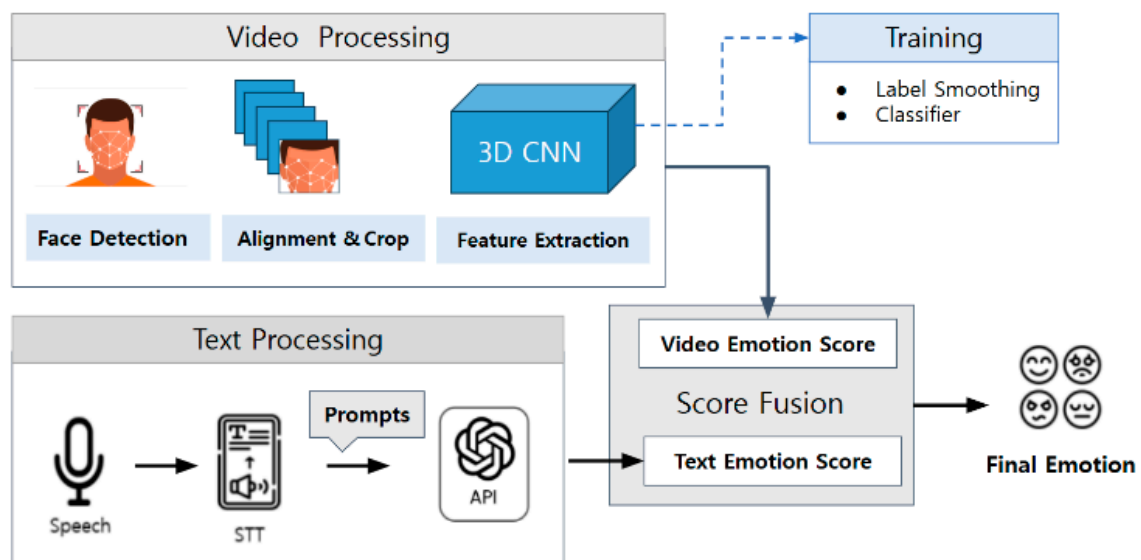
As shown in Table 1, existing approaches often rely on complex fusion mechanisms or resource-intensive models, which limits their applicability in real-time or low-resource scenarios. Our method addresses these limitations by adopting a simple yet effective decision-level fusion strategy.

Table 1. Summary of Recent Multimodal Emotion Recognition Approaches.

Study	Modalities	Fusion Strategy	Strengths	Limitations
Viegas et al. [13]	FAUs + Voice+ Skeleton	Feature-level + 2-stage (coarse-to-fine)	Video-level features improve performance over frame-level	No mention of real-time; rule-based stages may lack flexibility
Kumar et al. [14]	Audio + Video	Feature-level	Deep fusion with a meaningful combination of modalities	High-dimensionality and time-sync issues; limited generalization
Lee Dae-jong [15]	Still images + Video	Feature-level, Dynamic modeling	Effectively tracks emotional change over time	Inter-modality inconsistency can degrade performance
Jo and Jung [16]	Image + Text	Hybrid	Multidimensional emotion analysis (VAD)	Domain-specific text sensitivity, lack of context in real use
Moon [17]	Video + Voice	Feature-level	Specialized model for Korean data	Low generalization; mismatch between audio-video emotion info
Ortega et al. [18]	Audio + Video + Text	Intermediate Fusion	Balanced modality training, high performance	High data requirement; not suited for real-time inference
Yi et al. [19]	Text+ Audio	Intermediate + Late Fusion	Cross-modal attention improves context-sharing	High complexity, high computational cost
Li et al. [20]	EEG (single-channel)	Multi-Entropy Feature Fusion	Resource-efficient design using multi-entropy features	EEG-only approach

3. Materials and Methods

In this study, we propose an emotion recognition fusion method that applies a decision fusion method using text and video modalities based on a Korean video dataset, as shown in Figure 1.

**Figure 1.** Workflow of the proposed fusion framework.

The research flow is largely divided into the Text Information Processing section, which processes text data extracted from audio, and the Video Information Processing section, which processes facial image data. The detailed processing steps are divided into the step of extracting emotion scores from text, the step of implementing a model that learns videos to determine facial emotions, the step of extracting emotion scores using a video determination model, and finally the step of deriving the final emotion by fusing the text emotion scores and video emotion scores. At this time, emotions are based on seven basic emotions (happy, surprise, neutral, fear, disgust, angry, and sad).

3.1. Dataset

The video dataset is a dataset for emotion classification provided by AI-HUB [21], consisting of a total of 10,351 videos, and in each video, 100 actors perform lines in situations classified into seven basic emotions (happiness, surprise, blank expression, fear, disgust, anger, and sadness). Examining the data structure, it consists of 350 situational texts and utterances, videos of actors performing the utterances, and facial expression images for each emotion. The actors consisted of 43 males and 57 females, and their ages ranged from 20 to 40 s. Each video was produced in such a way that each person acted out the given utterance, given a situational text that could infer emotions. The video length varies from 2 to 4 s, depending on the speech duration, and consists of 90 to 150 frames. Figure 2 shows some frames from a video of a person acting out the line “Why are you saying this now?” in a situation where a person had planned to meet a friend but received a phone call 30 min before the appointment saying, “I don’t think I can make it today”.



Figure 2. Example of a sample: video labeled as “angry” with corresponding text, “Why are you telling me this now?”.

3.2. Text Emotion Score Extraction with GPT-4

One of the major challenges in emotion recognition research is the lack of high-quality, large-scale emotional datasets. To address this issue, recent studies have explored methods such as fine-tuning pre-trained models and using generative models to supplement emotional data.

In this study, we utilize a small-scale Korean emotion dataset consisting of approximately 350 text utterances. To compensate for data scarcity and obtain consistent emotion annotations, we send each utterance as a prompt to GPT-4 Turbo—an advanced large language model (LLM)—via the OpenAI API. The structure of the prompt and the inference process are detailed in Table 2. The model returns seven emotion scores per text. The overall process of emotion scoring using the LLM is illustrated in Figure 3. For consistency and reproducibility, we use a low temperature setting (0.2), which minimizes randomness in generation.

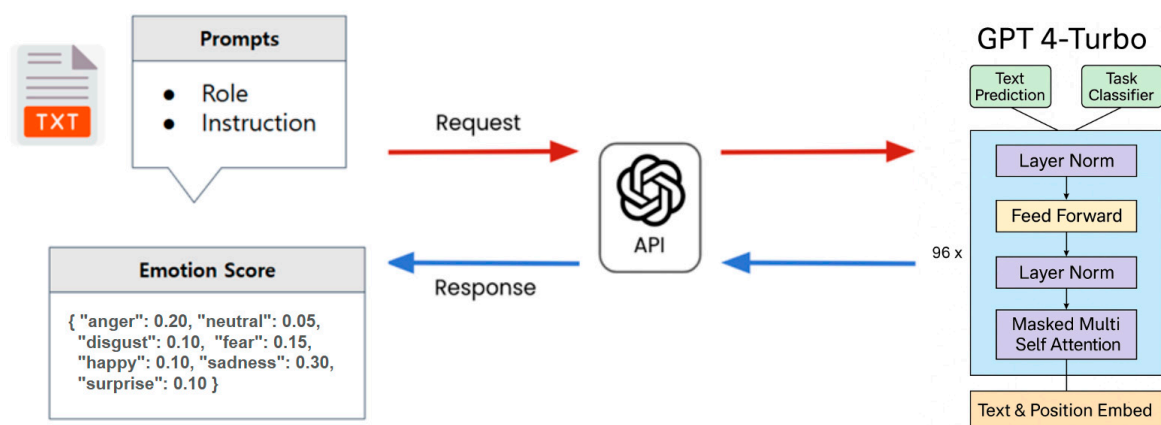


Figure 3. Text emotion scoring pipeline via large language model.

Table 2. Instruction prompt submitted to GPT-4 API.

Prompt Template and Inference Code
<pre> Prompt = '''Analyze the following text based on situation and assign a score from 0 to 1 for each of the following emotions: text = "{text}", situation = "{situation}", emotions = ['happy', 'surprise', 'neutral', 'fear', 'disgust', 'angry', 'sadness']. Return the result strictly as a Python Dictionary type with this structure: {{ "happy": <score>, "surprise": <score>, "neutral": <score>, "fear": <score>, "disgust": <score>, "angry": <score>, "sadness": <score> }}. Ensure the response is valid Dictionary only.''' response = client.chat.completions.create(model = "gpt-4-turbo", messages = [{"role": "system", "content": "You are a helpful assistant for sentiment analysis."}, {"role": "user", "content": prompt}], temperature = 0.2) </pre>

GPT-4 Turbo is based on the Transformer architecture and leverages a self-attention mechanism to understand contextual relationships. The input text is tokenized using Byte Pair Encoding (BPE), transformed into high-dimensional vectors via an embedding layer, and passed through multiple Transformer encoder layers to infer emotion scores in context.

While the use of GPT-4 Turbo involves external API calls, which may raise concerns about latency and computational cost, this approach serves as a proof-of-concept. It allows immediate deployment without requiring additional model training or dataset-specific fine-tuning. Furthermore, it enables access to the latest model updates and large-scale pre-learned emotional knowledge without incurring maintenance or retraining costs.

In practical applications, the framework can be adapted to run with lighter, locally fine-tuned models such as KoELECTRA or DistilBERT. These alternatives preserve the structure of our system while supporting faster inference and lower resource consumption. Therefore, our current implementation demonstrates conceptual validity while remaining flexible for future lightweight deployment.

3.3. Video Score Extraction with 3DCNN

Analyzing facial expressions in a video modality requires detecting and tracking faces in video frames. Three-dimensional convolutional neural networks (3DCNNs) are important in video-based emotion recognition. This algorithm improves emotion recognition performance by leveraging the inherent information in video data. Since video data have various lengths and resolutions, a preprocessing process is required to convert it into a certain format and normalize it to be used as input for the model.

First, the video frames are loaded, and the face is detected using 49 facial landmark information for each frame, and the face portion is accurately cropped. At this time, if the face is not detected, the frame is excluded. To maintain data consistency, we align the faces based on the positions of the two eyes and then reduce the size of the images to 240×240 . Afterwards, cropping is additionally performed based on the image center point to remove noise. Next, the frame size is adjusted to 128×128 to fit the designed 3D CNN model, and

16 consecutive frames are composed into one clip through clip unit division operation. At this time, if the video length is short enough to not fill the required number of clips, add blank clips filled with 0 and delete excess clips to keep the maximum number of clips to 6.

Figure 4 illustrates the architecture of the video emotion classification model, which is based on a compact 3D Convolutional Neural Network (3D CNN). The model takes input in the form of 16-frame RGB video clips (each of size $128 \times 128 \times 3$) and applies a $3 \times 3 \times 3$ convolution filter to extract spatio-temporal features that capture both local image patterns and temporal emotional dynamics. To ensure lightweight computation and real-time applicability, the model is designed with only two 3D convolutional layers, two 3D max-pooling layers, and a two-layer fully connected classifier. This efficient structure allows the model to maintain high performance while minimizing computational load. As confirmed by the model summary, it contains approximately 16.8 million trainable parameters, requires 814.74 million multiply–add operations (MACs) per inference, and has an estimated total memory footprint of 112.26 MB, including input, activation, and parameter sizes. Compared to more complex architectures such as Transformer-based multimodal models, the proposed video classifier offers a practical trade-off between performance and efficiency, making it suitable for deployment on edge devices and in real-time systems.

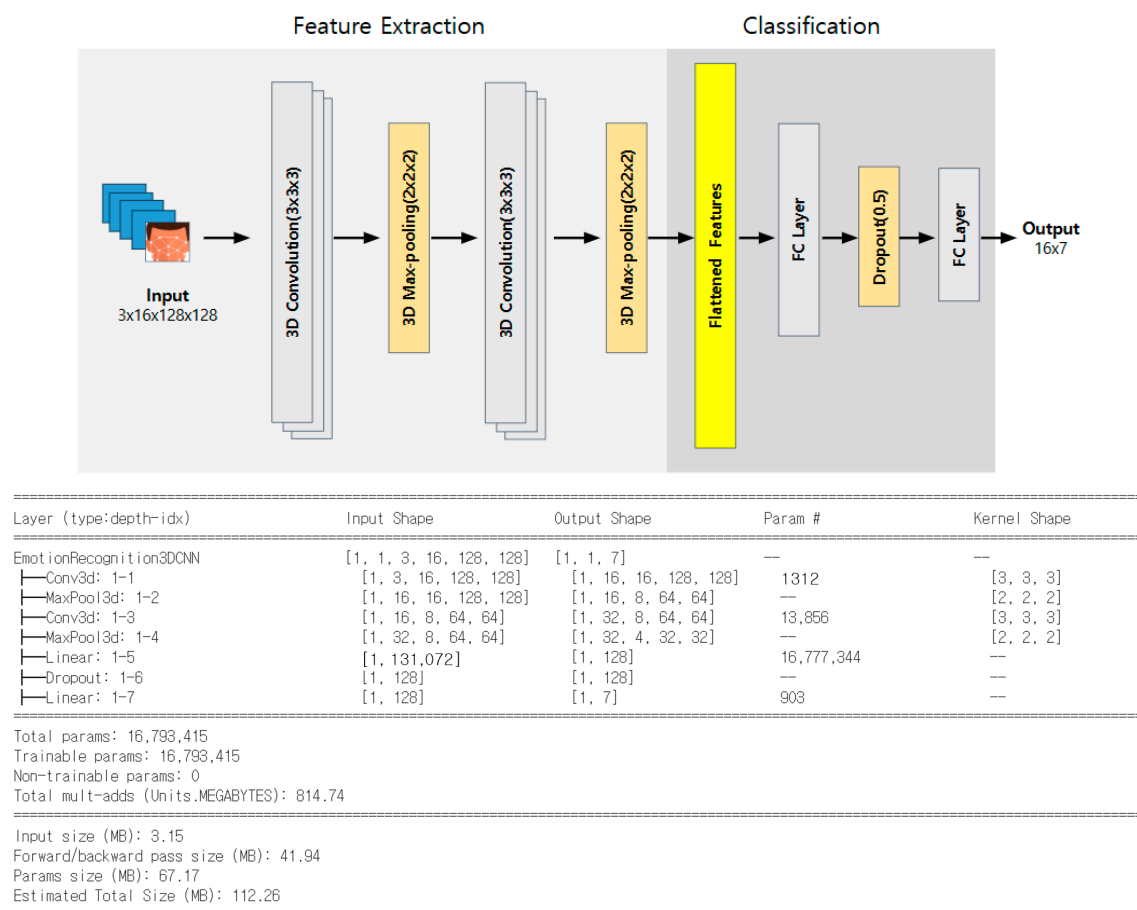


Figure 4. Model structure and specifications of the 3D CNN for video emotion analysis.

To avoid overfitting due to a small amount of data, we use the loss function of Equation (1) with a label-smoothing technique instead of dropout and a general cross-entropy loss

function. This equation represents the cross-entropy loss with label smoothing applied. C is the number of classes and α is the smoothing coefficient.

$$\mathcal{L} = -\sum_{i=1}^C \left((1 - \alpha) \cdot y_i^{one-hot} + \frac{\alpha}{C} \right) \cdot \log \sigma(z_i). \quad (1)$$

The optimizer uses Adam to apply an adaptive learning rate to improve learning speed. Additionally, the stability of emotion classification is improved by averaging multiple clips.

3.4. Score-Based Decision-Level Fusion

In emotion analysis, effectively combining multimodal data from text and video is essential for improving recognition accuracy. However, a simple average-based fusion method has limitations, as text and video convey emotional information in inherently different ways and exhibit varying levels of model reliability. To address this, we propose three fusion strategies (Method 1, Method 2, and Method 3) that integrate emotion scores from both modalities based on different theoretical perspectives on model confidence and inter-modality relationships.

Method 1 applies a fixed confidence-based weighting scheme to each modality. Specifically, the final emotion score is computed by weighting the text and video emotion scores using predefined confidence values (C_t for text, C_v for video), which are derived from each modality's standalone classification performance. This method assumes that model confidence remains stable across all samples, providing a consistent and interpretable baseline.

$$w_t = \frac{C_t}{C_t + C_v}, w_v = \frac{C_v}{C_t + C_v} \quad S_{final} = w_t S_{text} + w_v S_{video}. \quad (2)$$

Method 2, in contrast, introduces a dynamic weighting mechanism based on the correlation between the emotion scores of the two modalities. If the two scores are strongly correlated, the method simply sums them; if they are weakly correlated, it falls back to the confidence-based weighting from Method 1. This method assumes that real-time agreement between modalities is an indicator of reliability and thus adapts fusion weights accordingly.

These two methods are theoretically distinct: Method 1 is static and performance-driven, whereas Method 2 is adaptive and context-sensitive, leveraging inter-modal consistency. The comparison between them allows us to evaluate the effectiveness of static vs. dynamic fusion strategies in multimodal emotion recognition.

$$NCC(S_{text}, S_{video}) = \frac{\sum (S_{text} - \mu_t)(S_{video} - \mu_v)}{\sqrt{\sum (S_{text} - \mu_t)^2} \sqrt{\sum (S_{video} - \mu_v)^2}}. \quad (3)$$

$$S_{final} = \begin{cases} S_{text} + S_{video}, & \text{if } NCC > r \\ w_t S_{text} + w_v S_{video}, & \text{otherwise} \end{cases}$$

Method 3 combines emotion scores from text and video modalities using Dempster-Shafer Theory (DST), a framework for reasoning under uncertainty. While DST has been used in prior multimodal fusion studies, our method introduces two key adaptations tailored for emotion recognition tasks with soft probabilistic scores. First, we convert soft emotion scores into basic belief assignments (BBAs), enabling the model to handle ambiguous emotional states. Second, we incorporate modality-specific confidence into the DST combination rule, allowing more reliable sources to influence the final decision. These adaptations enhance DST's applicability to emotion recognition in uncertain and

low-resource environments. Unlike conventional DST applications that rely on discrete class labels, our approach enables graded reasoning over soft emotional probabilities.

$$m(Z) = \frac{\sum_{X \cap Y = A} m_{\text{text}}(X) \cdot m_{\text{video}}(Y)}{1 - \sum_{X \cap Y = \emptyset} m_{\text{text}}(X) \cdot m_{\text{video}}(Y)} \quad m_{\text{text}}(X) = S_{\text{text}}, \quad m_{\text{video}} = S_{\text{video}} \quad (4)$$

4. Results

4.1. Text Emotion Recognition Performance

Emotion classification for text was performed on 350 data, each labeled with 50 emotions per emotion, using the ChatGPT API (accessed via <https://platform.openai.com>, GPT-4 Turbo, version as of April 2025), a large language model (LLM), to classify the seven basic emotions. The results are shown in Table 3.

Table 3. Performance of the text modality in emotion recognition.

	Accuracy	Precision	Recall	F1-Score
Happy	0.82	0.66	0.82	0.73
Surprise	0.68	0.51	0.68	0.58
Neutral	0.78	0.50	0.78	0.61
Fear	0.38	0.63	0.38	0.47
Disgust	0.20	0.91	0.20	0.33
Angry	0.74	0.56	0.74	0.64
Sad	0.46	0.64	0.46	0.53
Average	0.58	0.63	0.58	0.56

Table 4 compares the results of emotion classification on text when only simple utterances are given and when contextual information is given along with the utterances. Figure 5 shows the results as a confusion matrix. In simple utterances, it is particularly difficult to distinguish between fear, disgust, anger, and sadness. When context information is provided, it can be confirmed that the performance improves.

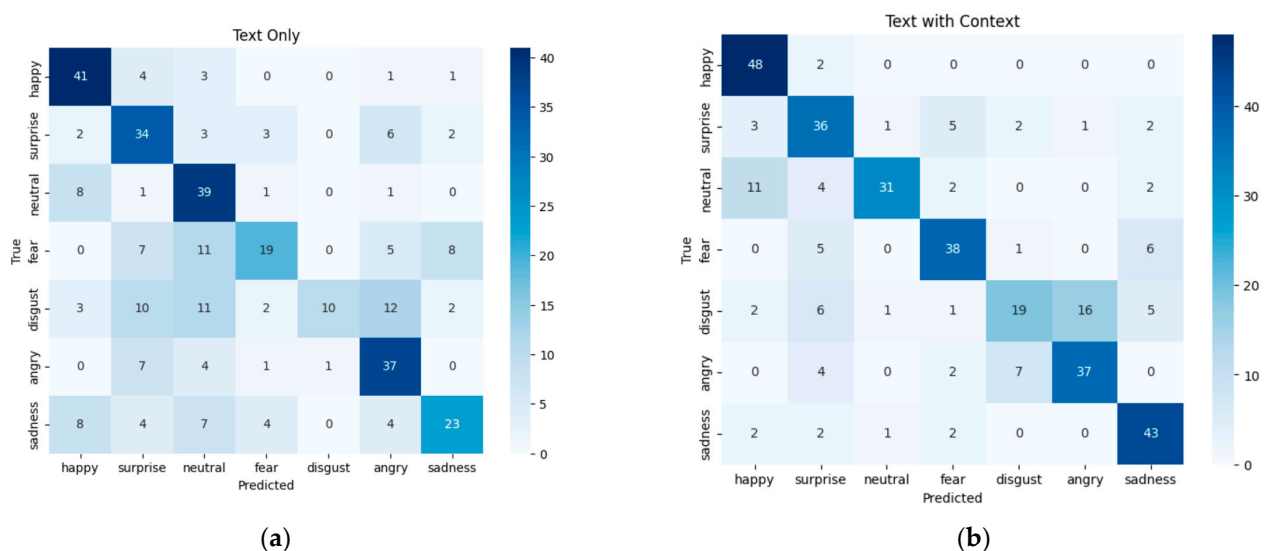


Figure 5. Confusion matrix: (a) Text Only: when emotions are recognized by considering only the spoken text performed by the actor. (b) Text with Context: when emotions are recognized by considering both the speech situation and the speech text performed by the actor.

Table 4. Comparison of emotion recognition performance: Text Only vs. Text with Context.

Emotion	Text Only	Text with Context
Happy	0.82	0.96
Surprise	0.68	0.72
Neutral	0.78	0.62
Fear	0.38	0.76
Disgust	0.20	0.38
Angry	0.74	0.74
Sad	0.46	0.86
Average	0.58	0.72

4.2. Video Emotion Recognition Performance

For training, only 4900 video samples, which are a subset of the entire dataset, were used. These samples were divided into three groups—training, validation, and testing—as shown in Table 5, and experiments were conducted accordingly.

Table 5. Composition and size of the experimental dataset.

	Train	Validation	Test
Dataset	3430	735	735

Table 6 presents the training parameters: each clip consists of 16 frames, each video is divided into 6 standard clips, the input frame size is adjusted to 128×128 , and a single filter with a size of $3 \times 3 \times 3$ is applied.

Table 6. Hyperparameter settings for model training.

Parameters	Values
Frames per clip	16
Standard clips per video	6
Input frame size	128×128
Filter size	$3 \times 3 \times 3$
Batch size	8
Optimizer	Adam
Initial learning rate	0.001
Number of epochs	50

As a result, Figure 6 is a graph showing the changes in loss and accuracy during the model's training process, and the classification accuracy was found to be approximately 74%. (a) In the graph, the loss decreases quickly and then stabilizes, but the validation loss decreases and then fluctuates at a certain level. (b) In the graph, the learning accuracy increases continuously and approaches 100%, while the verification accuracy initially increases and then remains at a relatively low level, showing a tendency to oscillate. This situation indicates overfitting due to insufficient data, and data augmentation is needed to address the issue. In this study, however, we assumed a low-resource scenario and employed early stopping to save and use the best-performing model during training.

As shown in Figure 7 and Table 7, the video-based emotion classification model outperformed the text-only emotion classification model overall. Nevertheless, distinguishing between *fear*, *disgust*, and *anger* remains challenging.

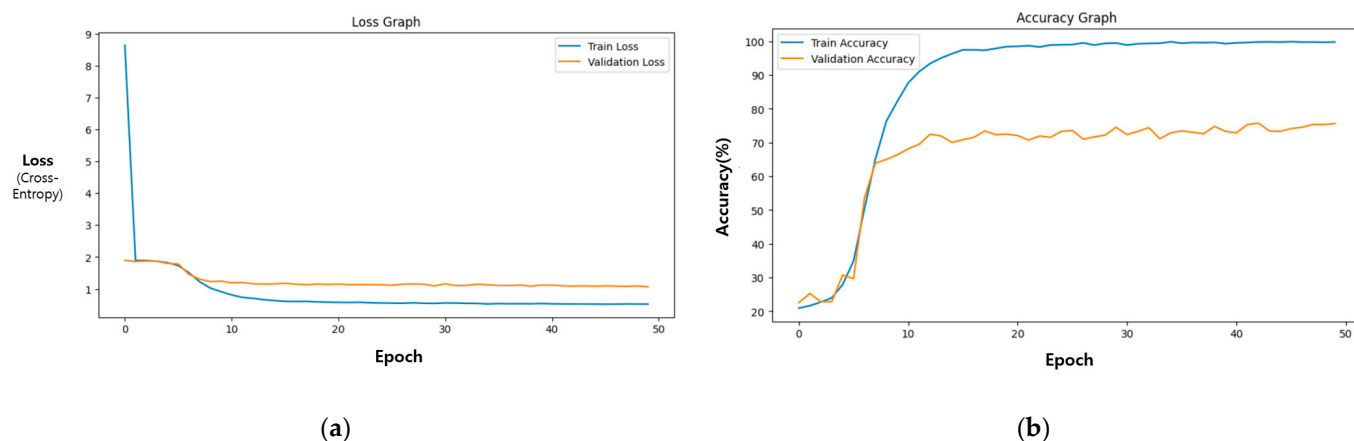


Figure 6. Training and validation performance across epochs: (a) training and validation loss curves over epochs; (b) training and validation accuracy curves over epochs (accuracy shown as percentage).

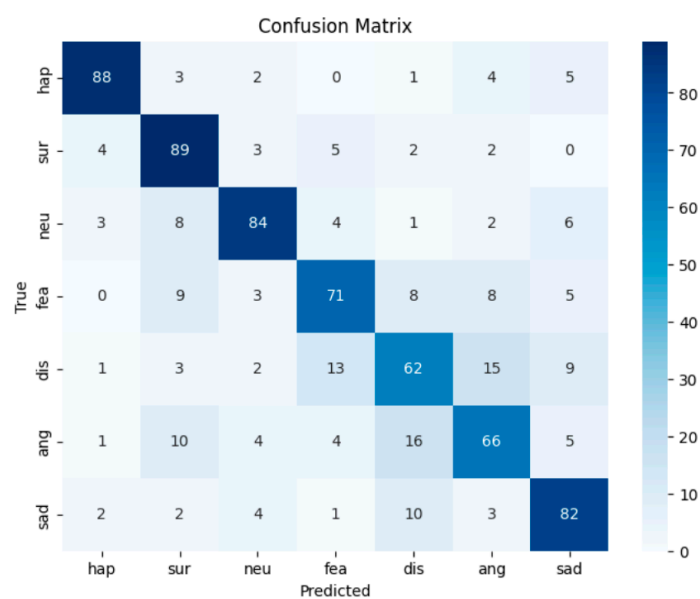


Figure 7. Confusion matrix of the video-based emotion recognition model.

Table 7. Performance of the video modality in emotion recognition.

	Accuracy	Precision	Recall	F1-Score
Happy	0.85	0.89	0.85	0.87
Surprise	0.85	0.72	0.85	0.78
Neutral	0.78	0.82	0.78	0.80
Fear	0.68	0.72	0.68	0.70
Disgust	0.59	0.62	0.59	0.60
Angry	0.62	0.66	0.62	0.64
Sad	0.79	0.73	0.79	0.76
Average	0.74	0.74	0.74	0.74

4.3. Score-Based Fusion Performance

The results in Table 8 are the result of applying Method 1, which calculates a weighted average of the video and text emotion scores reflecting the reliability of each model and then determines the final emotion. In other words, it is a fusion method that dynamically adjusts the weights reflecting the reliability of each data modality.

Table 8. Performance of Method 1: fixed confidence-based fusion.

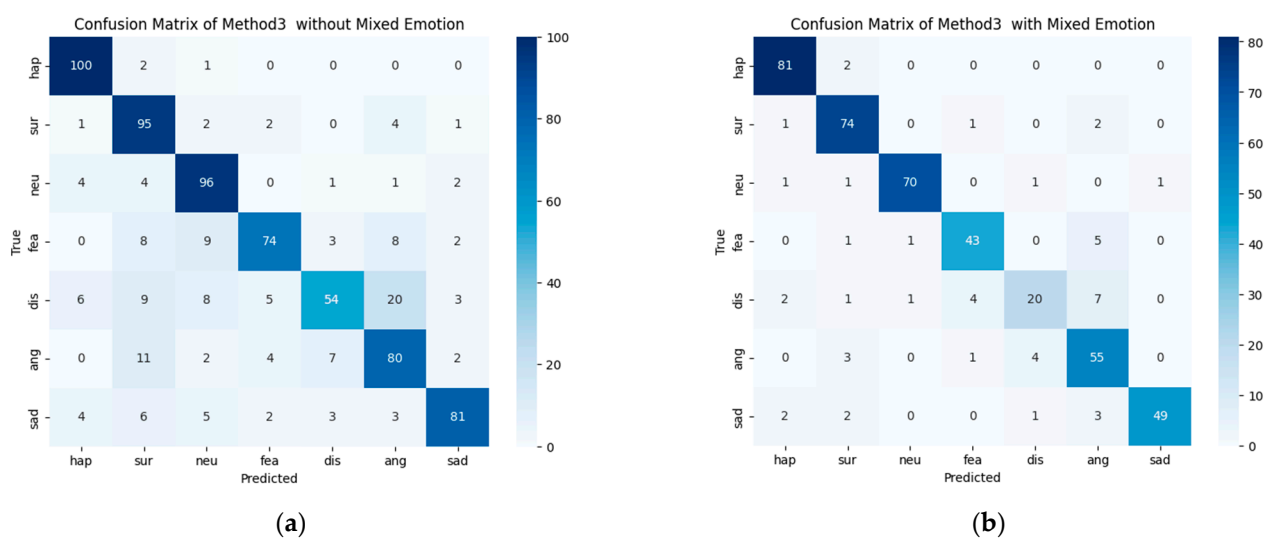
	Accuracy	Precision	Recall	F1-Score
Happy	0.97	0.87	0.97	0.92
Surprise	0.90	0.70	0.90	0.79
Neutral	0.89	0.78	0.89	0.83
Fear	0.71	0.85	0.71	0.77
Disgust	0.51	0.79	0.51	0.62
Angry	0.75	0.69	0.75	0.72
Sad	0.78	0.89	0.78	0.83
Average	0.80	0.80	0.79	0.79

Table 9 shows the results of an algorithm that calculates the correlation between the text and video emotion scores and then dynamically adjusts the weights according to the similarity to determine the final emotion. Unlike Method 1, this method uses the Normalized Correlation Coefficient (NCC) to determine whether the emotion distributions of the text and video are similar.

Table 9. Performance of Method 2: correlation-based fusion.

	Accuracy	Precision	Recall	F1-Score
Happy	0.97	0.87	0.97	0.92
Surprise	0.90	0.70	0.90	0.79
Neutral	0.89	0.79	0.89	0.84
Fear	0.71	0.84	0.71	0.77
Disgust	0.50	0.79	0.50	0.61
Angry	0.75	0.69	0.75	0.72
Sad	0.78	0.87	0.78	0.82
Average	0.80	0.79	0.79	0.79

Figure 8 shows the results of applying Method 3, which combines text and video emotion scores by modifying the Dempster–Shafer Theory (DST) and then determines the final emotion: (a) is the case where compound emotion is not considered, and (b) is the case where compound emotion is considered.

**Figure 8.** Performance comparison of Method 3: (a) without compound emotions; (b) with compound emotions.

When using compound emotions, prediction became more challenging compared to the conventional single-emotion model. However, strong emotions such as anger, disgust, and fear were expressed in a more complex manner. Nonetheless, as seen in the sample data from Table 10, compound emotions better captured the contextual nuances of the situation.

Table 10. Examples of compound emotion usage in Method 3.

Item	Content
Text	You shouldn't buy that car.
Situation Scenario	A breaking news alert suddenly appeared on the peaceful TV screen. The report stated that an SUV driving on the highway had rolled over into the adjacent lane, but the airbags failed to deploy, resulting in the tragic death of the entire family.
True Label	disgust
Predict Label	Fear_disgust

Figure 9 shows a comparison of performance between single-modality models and the proposed multi-modal fusion approaches in terms of recall and F1-score. The analysis revealed that the proposed methods consistently achieved superior performance across most emotion categories. As summarized in Table 11, the proposed fusion method achieved higher accuracy (0.80) and F1-score (0.79) compared to the text-only and video-only baselines. Since accurately identifying negative emotions is particularly important, recall was emphasized as a key evaluation metric in this study.

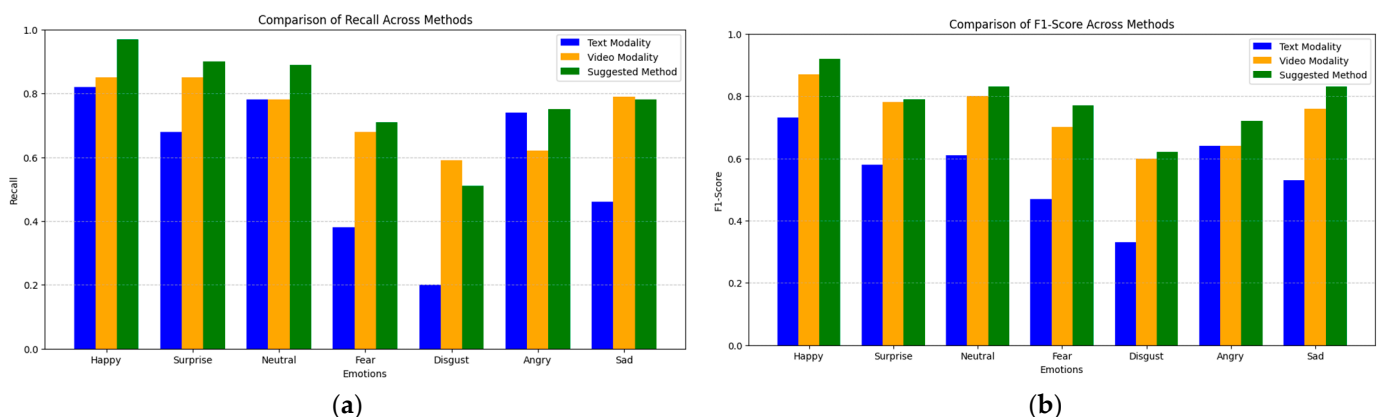


Figure 9. Comparison of fusion methods across emotion classes: (a) recall; (b) F1-score.

Table 11. Performance comparison of the proposed fusion method with single-modality baselines.

Research	Accuracy	F1-Score
Text Modality	0.58	0.56
Video Modality	0.74	0.74
Proposed Method	0.80	0.79

Furthermore, repeated experiments demonstrated that the performance of the three fusion strategies was stable, with minimal variation across runs. Although minor differences were observed in emotion-specific metrics—especially in harder-to-classify emotions like disgust—the macro-average F1-score difference between methods was within 1%, suggesting that all proposed methods are comparably effective overall.

Table 12 compares the performance with existing studies based on the Korean video dataset. As shown in Table 9, the proposed method achieved the highest performance with an accuracy of 0.80 and an F1-score of 0.79, outperforming existing approaches such as

those by Chanyoung J [16] and Seokho [17]. This demonstrates the effectiveness of the proposed fusion strategy over traditional multimodal methods.

Table 12. Comparison of proposed method with existing studies.

Research	Accuracy	F1-Score
Text + Video-Frame-avg [22]	0.65	0.65
Chanyoung J [16]	0.74	0.74
Seokho [17]	0.69	0.70
Proposed Method	0.80	0.79

5. Discussion

A key challenge in emotion recognition research is the scarcity of large-scale annotated datasets, particularly for the Korean language. This limitation reduces model generalization and makes feature-level fusion difficult due to high-dimensional and asynchronous data. To address this issue, we leveraged pre-trained large-scale language models and proposed an emotion score-based decision fusion method that combines text and video information, aiming to complement the limitations of single-modality approaches while reducing learning complexity.

For the text modality, emotion scores were obtained by submitting text prompts to GPT-4 Turbo, a large language model (LLM), via the OpenAI API. The model returned a vector of seven emotion scores per utterance. In simple expressions, the model achieved 58% accuracy, which improved to 72% when contextual information was considered, demonstrating its effectiveness in capturing nuanced emotional cues. However, it still showed difficulty in accurately classifying negative emotions such as fear, disgust, and sadness.

For the video modality, we trained a model based on a 3D convolutional neural network, which achieved strong performance (74% accuracy) despite the limited dataset. The proposed 3D CNN model was designed to be lightweight, containing only two convolutional layers, two pooling layers, and two fully connected layers, with approximately 16.8 million parameters and a total memory footprint of around 112 MB. This compact design enables efficient inference suitable for edge computing.

To further enhance overall classification, we implemented and compared three fusion algorithms based on the reliability of each modality: (1) confidence-based weighting, (2) correlation-aware adjustment, and (3) Dempster–Shafer Theory (DST)-based combination. These methods yielded comparable performance overall, with minor variations in the classification of disgust. Compared to previous studies using Korean datasets, our approach achieved a favorable trade-off between model complexity and recognition performance.

Our approach maintains lightweight design principles at every level: (i) text-based emotion recognition is conducted using LLM APIs during inference only, without local fine-tuning or additional training; (ii) the video modality uses a simple 3D CNN model with no complex temporal attention or Transformer blocks; and (iii) fusion is performed using score-level decision methods rather than neural-level integration, minimizing additional computational overhead. These design choices result in a multimodal emotion recognition framework that is both accurate and highly suitable for real-time applications.

Although our proposed fusion methods improved the overall classification performance, the emotion class disgust consistently exhibited low recognition accuracy. This may be attributed to its semantic and expressive proximity to Anger, leading to frequent confusion between the two. To address this, we considered merging closely related negative emotions, such as anger and disgust, into a single broader category—a strategy supported in prior affective computing literature. While this could improve classification robustness and mitigate class imbalance, it would also reduce emotional granularity and limit interpretive resolution.

Thus, we chose to maintain separate classes in this study, but we recognize the value of category merging as a promising direction for future work, especially in real-time or low-resource scenarios. In addition, we aim to incorporate audio as a supplementary modality and explore the use of combined emotion categories to gain a deeper understanding of affective states. Ultimately, our goal is to develop a lightweight, real-time multimodal emotion recognition system that balances accuracy with practical deployability.

Author Contributions: Methodology, E.-H.K.; Formal analysis, M.-J.L.; Investigation, M.-J.L.; Writing—original draft, E.-H.K.; Writing—review & editing, J.-H.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2023R1A2C1006149).

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Seo, E.; Koo, S.J.; Kim, Y.J.; Min, J.E.; Park, H.Y.; Bang, M.; Lee, E.; An, S.K. Reading the Mind in the Eyes Test: Relationship with Neurocognition and Facial Emotion Recognition in Non-Clinical Youths. *Psychiatry Investig.* **2020**, *17*, 835–839. [[CrossRef](#)] [[PubMed](#)]
2. Kossaifi, J.; Toisoul, A.; Bulat, A.; Panagakis, Y.; Hospedales, T.; Pantic, M. Factorized higher-order cnns with an application to spatio-temporal emotion estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6060–6069.
3. Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J.G. Emotion recognition in human-computer interaction. *IEEE Signal Process. Mag.* **2001**, *18*, 32–80. [[CrossRef](#)]
4. Zadeh, A.B.; Liang, P.P.; Poria, S.; Cambria, E.; Morency, L.P. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 2236–2246.
5. Kalateh, S.; Estrada-Jimenez, L.A.; Nikghadam-Hojjati, S.; Barata, J. A systematic review on multimodal emotion recognition: Building blocks, current state, applications, and challenges. *IEEE Access* **2024**, *12*, 103976–104019. [[CrossRef](#)]
6. Njoku, J.N.; Caliwag, A.; Lim, W.; Kim, S.; Hwang, H.-J.; Jeong, J.-W. Deep learning based data fusion methods for multimodal emotion recognition. *J. Korean Inst. Commun. Inf. Sci.* **2022**, *47*, 79–87. [[CrossRef](#)]
7. Ramachandram, D.; Taylor, G.W. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Process. Mag.* **2017**, *34*, 96–108. [[CrossRef](#)]
8. Lucía, M.J.; Revuelta, P.; García, Á.; Ruiz, B.; Vergaz, R.; Cerdán, V.; Ortiz, T. Vibrotactile captioning of musical effects in audio-visual media as an alternative for deaf and hard of hearing People: An EEG study. *IEEE Access* **2020**, *8*, 190873–190881. [[CrossRef](#)]
9. Zhang, H. Expression-EEG based collaborative multimodal emotion recognition using deep autoencoder. *IEEE Access* **2020**, *8*, 164130–164143. [[CrossRef](#)]
10. Cimtay, Y.; Ekmekcioglu, E.; Caglar-Ozhan, S. Cross-subject multimodal emotion recognition based on hybrid fusion. *IEEE Access* **2020**, *8*, 168865–168878. [[CrossRef](#)]
11. Sahu, G.; Vechtomova, O. Adaptive fusion techniques for multimodal data. *arXiv* **2019**, arXiv:1911.03821.
12. Palladin, E.; Dietze, R.; Narayanan, P.; Bijelic, M.; Heide, F. Samfusion: Sensor-adaptive multimodal fusion for 3d object detection in adverse weather. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; Springer Nature: Cham, Switzerland, 2024; pp. 484–503.
13. Viegas, C. Two stage emotion recognition using frame-level and video-level features. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, 16–20 November 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 912–915.
14. Kumar, H.; Martin, A. Audio-Video Based Fusion Mechanism Using Deep Learning for Categorical Emotion Recognition. In Proceedings of the 2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS), Erode, India, 18–20 October 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 595–602.
15. Lee, D.-J.; Go, H.; Chun, M. Facial emotion recognition using a fusion model of still images and video. *J. Korean Inst. Intell. Syst.* **2005**, *15*, 573–580.
16. Cho, C.-Y.; Jeong, H.-J. Multimodal emotion recognition system using facial video and multidimensional emotion-based text. *J. Korea Inf. Technol. Serv.* **2023**, *21*, 39–47.

17. Moon, S.; Kim, S.B. Multimodal Deep Learning Model for Korean Video emotion Classification. In Proceedings of the Autumn Conference of the Korean Institute of Industrial Engineers, Seoul, Republic of Korea, 13 November 2020; pp. 2944–2955.
18. Ortega, J.D.; Senoussaoui, M.; Granger, E.; Pedersoli, M.; Cardinal, P.; Koerich, A.L. Multimodal fusion with deep neural networks for audio-video emotion recognition. *arXiv* **2019**, arXiv:1907.03196.
19. Yi, M.H.; Kwak, K.C.; Shin, J.H. HyFusER: Hybrid Multimodal Transformer for Emotion Recognition Using Dual Cross Modal Attention. *Appl. Sci.* **2025**, *15*, 1053. [[CrossRef](#)]
20. Li, J.; Feng, G.; Ling, C.; Ren, X.; Liu, X.; Zhang, S.; Wang, L.; Chen, Y.; Zeng, X.; Chen, R. A Resource-Efficient Multi-Entropy Fusion Method and Its Application for EEG-Based Emotion Recognition. *Entropy* **2025**, *27*, 96. [[CrossRef](#)] [[PubMed](#)]
21. Emotion Classification Dataset. 2018. Available online: <https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=259> (accessed on 2 February 2025).
22. Kim, E.-H.; Shin, J.-H. Multimodal emotion recognition from speech videos using a pre-trained model. *J. Smart Media* **2024**, *13*, 19–27.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.