Advanced Optimal Cross-Modal Fusion Mechanism for Audio-Video Based Artificial Emotion Recognition

Himanshu Kumar*, Martin Aruldoss
Department of Computer Science, Central University of Tamil Nadu, Thiruvarur, India.
E-mail: himanshukphd20@students.cutn.ac.in, martin@cutn.ac.in[2]
*Corresponding author

Keywords: multimodal fusion, cross-modal fusion, emotion recognition, artificial emotion intelligence, fusion mechanism

Received: October 22, 2024

The advanced technology of artificial emotional intelligence has greatly contributed to multimodal emption recognition task. Emotion recognition has played a crucial role in many domains, like communication, e-learning, mental healthcare, contextual awareness, and customer satisfaction. As real-time data continues to expand, addressing the problem of emotion recognition has become critical and complex. A key challenge lies in recognizing emotions from multimodal heterogeneous input sources, aligning extracted features, and developing robust emotion recognition models. In this study, we explore a cross-modal (audio and video modality) fusion mechanism for emotion recognition, effectively addressing the associated feature complexities. We have used 2D-CNN and 3D-CNN deep learning models for audio and video feature extractions and developed robust models for emotion recognition. This study emphasizes the importance of Compact Bilinear Gated Pooling (CBGP) cross-modal fusion mechanism and highlights the contribution of fusing the features from audio and video modalities for emotion recognition. It also discusses the working principle and comparison performance with other peer cross-modal fusion techniques such as FBP and CBP. The performance of advanced cross-modal fusion is compared to baseline traditional cross-modal fusion mechanisms including EF-LSTM, LF-LSTM, Graph-MFN, hybrid fusion and transformer model based fusion mechanisms such as, attention fusion and transformer fusion. This experiment is performed on benchmark datasets CMU-MOSEI and achieves an accuracy of 80.3%, F1-score of 79.2%, and MAE of 54.2%.

Povzetek: Predstavljen je napredni mehanizem optimalne fuzije med modalnostmi za umetno prepoznavanje čustev na podlagi avdio-video posnetkov. Študija uporablja 2D- in 3D-CNN za ekstrakcijo značilnosti, poudarja pomen CBGP fuzije in dosega odlične razultate na naboru podatkov CMU-MOSEI.

1 Introduction

Emotion recognition is being successfully used in many domains and applications. The adoption of this technology has grown rapidly in healthcare, e-learning and advertising [1]. Initially, emotion recognition was limited to with unimodal approaches, but it has now gained more popularity with the advancement of multimodal approaches and enhanced techniques. Its growing demand has expanded the scope for exploring various directions of research in emotion recognition. Multimodal data inherently contains rich information and has the potential to learn meaningful patterns from extracted features. In our study, we intend to achieve emotion recognition by combining features extracted from audio and video modalities and employing a fusion mechanism. This study explores the cross-modal fusion approach, where the term ‘cross modal fusion’ refers to integrating essential features from heterogeneous input sources, further this integration helps in training deep learning models and classifying emotions effectively. Advanced cross-modal fusion mechanisms are categorized in three types: Factorized bilinear pooling (FBP) [2], Compact bilinear pooling (CBP) [3], and Compact Bilinear Gated Pooling (CBGP) [4].

Emotion recognition from audio and video modalities are very crucial because audio and video (collection of image frames) gives a wide range of information regarding, pitch, tone, image texture, facial movements, and facial expressions [5]. To train a model it is easy to extract features within the same modality and from another modality. This type of feature extraction leads to training a deep learning model to fine grained emotion classification tasks [6]. To work with different modalities, the most important and primary step is to extract the features from both the modalities. After preprocessing and cleaning the features, it is required to align those features, and combine only those features which have essential information and can help to train a deep learning model [7]. This study uses two different deep learning models, one is 2D-CNN [8] for audio modality and other is 3D-CNN [9] for video modality. As per the previous studies, this study aims to explore the advanced fusion mechanism such as Factorized bilinear pooling (FBP), Compact bilinear pooling (CBP), and Compact Bilinear Gated Pooling (CBGP). This study compares the advanced fusion approaches with state-of-the-art fusion approaches such as early fusion, late fusion, and hybrid fusion, as well as transformer model based fusion techniques such as attention fusion and transformer fusion.

The research contribution of the proposed work are as follows:

- Highlights the limitations of traditional fusion mechanisms, such as high dimensionality, suboptimal interdependency modeling, and challenges in fine-grained emotion classification.
- Addresses a critical gap to reduce the computational errors and improve the sustainability of audio-video emotion recognition systems.
- Introduces a novel gating unit and cross-modal fusion approach using factorized bilinear pooling and compact bilinear pooling, addressing the inefficiencies in traditional fusion methods. This solution enhances feature interaction and reduces computational complexity.
- Employs lightweight 2D-CNN and 3D-CNN architectures for audio and video modalities, respectively, avoiding the need for pruning and quantization while maintaining network simplicity.
- This design minimizes computational overhead associated with insignificant weights and neurons. Validates the model's accuracy and compares the performance of all three advanced cross-modal fusion mechanisms using the benchmark dataset CMU-MOSEI.
- Validates the model's accuracy and compares the performance with baseline, and traditional state-of-the-art fusion approaches: early fusion, late fusion and hybrid fusion.
- Comprehensive discussion with transformer model based fusion approaches: attention fusion and transformer fusion.
- The proposed approach ensures scalability and sustainability, contributing to the development of more resource-efficient deep learning models for real-world applications.

The rest of the paper is organized as follows: section 2 reviews the literature on feature extraction and traditional fusion mechanism and highlights the related work and research gap. Section 3 introduces the advanced cross-modal fusion approaches. Section 4 presents the training model and experimental setup, section 5 provides the result and discussion, and finally, Section 6 concludes the paper and suggests future scope.

2 Literature review

This section offers an overview of the features of audio-video modalities, and the existing fusion mechanisms in multimodal emotion recognition, along with a detailed review. Table 1 summarizes the related work and some baseline cross-modal fusion mechanisms, particularly for emotion recognition in audio-video modalities using the CMU-MOSEI dataset.

2.1 Feature extraction

Before feature extraction, the raw input dataset is pre-processed to ensure it is free from noise, missing values and other inconsistencies [10]. Feature extraction is a crucial part of feature engineering in any classification model, which yields critical information from the input data. Feature sets act as input vectors for a deep learning model, containing all the necessary information about the modalities that help the model learn patterns [11]. This section reviews the features and feature sets of audio and video modalities utilized in previous research studies.

i. Audio features

To effectively train deep learning models with audio features, feature extraction tools and libraries such as LibROSA [12], OpenSMILE [13], and pyAudioAnalysis [14] has proven indispensable. These tools are essential to process and extract the meaningful features, offering a robust foundation for building a deep learning model. The process begins with raw audio data undergoing a preprocessing step. After preprocessing, audio features are extracted using these tools and libraries. These features contain the information about acoustic properties [15] of audio utterances embedded within the video track. The extracted feature provides crucial information about various feature segments such as pitch, tone, energy, rhythm, and spectral attributes [16]. These properties capture many useful insights from raw audio data to train the deep learning model, which drives to classify the emotional state. Some most widely used extracted key features include:

- Mel-Frequency Cepstral Coefficients (MFCCs) [17]: Derived from spectrograms to represent the audio signal in a form humans perceive.
- Spectral features [18]: Attributes such as spectral centroid, roll-off, bandwidth that highlight energy distribution across frequencies.
- Variations in pitch, frequencies, amplitude [19]: Capturing changes in voice that are indicative of different emotions.
- Energy and intensity levels [19]: it represents the changes in signal strength, where low intensity often refers to 'sad' and high intensity correlates with 'excitement or happy' emotions.

Video features

Extracting video features is an essential step to train a deep learning model for emotion recognition. This process takes multiple sub-steps like extracting frames from the video, setting the frames per second, and extracting per frame features. After extracting frames, it is required to preprocess the entire frames as per standards for emotion recognition.

This preprocessing includes tasks such as frame sampling [20], facial feature alignment [21], discarding irrelevant frames and reducing variations.

Previous studies have explored two broad approaches to extracting the features from frames: appearance-based features and geometric-based features [22].

Appearance-based features: These features describe the visual characteristics as features of a picture within a specific frame, such as the face, facial expression, expression textures, sharpness, and facial movements [23]. These features provide pure cues and essential information for recognizing emotions.
Geometric-based features: These features are determined based on the calculation of facial landmarks, jaw movements, eyebrow movements, expression coordinates, relative positions, distance, arcs, shape angles, texture angles, and other facial action parameters [24].

These features are extracted using machine learning algorithms [25]–[28], traditional feature extraction techniques [29]–[32], and currently deep neural network models [12], [33]–[35]. Python libraries and frameworks are now widely used for feature extraction processes, enabling the development of more robust models for emotion recognition.

2.2 Feature fusion mechanism

After extracting features from both the audio and video modalities, an integration process is required to combine them effectively. This process, known as information fusion or feature fusion, involves aligning the key features from each modality obtained during the feature extraction and fusing them into a unified representation [36]. The goal is to synchronize the features of both modalities to collaboratively recognize emotions with higher accuracy. In this fusion process, the integrated features are first used to train a deep learning model. The model is then validated to ensure its accuracy and reliability in emotion recognition.

Early fusion

Early fusion [5] is one of the simplest and most fundamental mechanisms for multimodal fusion. In this fusion mechanism, features from different modalities are first aligned and integrated after extraction and then fed into a deep neural network model as input. This method combines audio and features into a single unifies feature vector, by applying the concatenation or elementwise operations such as addition, multiplication the, processed by a deep learning classification model for emotion recognition.

Late fusion

To address the limitations of early fusion, another basic fusion mechanism, late fusion [37], was introduced. A significant amount of research has shifted towards this fusion mechanism to develop more robust emotion classification models. In late fusion, each modality is first pre-processed, analyzed, and fed into a deep neural network model as input. The outputs from these classification models are then combined at a later stage. The advantage of this fusion mechanism lies in its ability to fuse features with low dimensionality and accurately classify emotions.

Hybrid fusion

Hybrid fusion [38] is hybridization of early and late fusion, integrating the feature properties of both fusion principles. It is considered superior to early and hybrid fusion in emotion classification. This fusion is particularly useful for addressing the challenges associated with the complexity of early and late fusion. Hybrid fusion can be applied in two phases; first, during the initial feature interaction, and second, after the model has been trained. However, this fusion technique fails to manage large parameters and complex features, where extracting and combining correlation based spatiotemporal feature information and identifying patterns are critical in multimodal emotion recognition. Hence, hybrid fusion needs further improvements to deal with complex multimodal datasets.

Attention fusion

Attention fusion [39] is a mechanism that focuses on fusing only the most relevant and crucial features after extracting all the features and generating feature maps from multimodal inputs. The advantage of this approach is to excel in handling both inter-modality and intra-modality interactions effectively. However, a major drawback of this fusion mechanism arises when feature alignment errors occur in spatiotemporal datasets or when sequence synchronization is lacking. Such issues lead to weak attention scores, increasing data complexity and computational burden [40]. There are two types of attention fusion mechanisms: self-attention [41] and multi-head attention [13]. Self-attention fusion sequentially captures interactions within a single modality, while multi-head attention focuses on every aspect of feature representation and captures interactions as output from multiple heads in parallel.

Transformer fusion

Transformer fusion [42] is an advanced approach of fusion mechanism that leverages pre-trained transformer models, which scales well on long sequencing data due to their ability to perform parallel computations. This fusion approach is particularly suitable for text-based emotion recognition tasks and natural language processing (NLP) applications, as it processes all token embeddings simultaneously. However, transformer fusion is less efficient when applied with audio and video modalities together. This limitation arises from the tokenization-synchronization trade-off between audio and video frame intervals, and positional embedding segments can lead to a loss of critical information and feature correlations in these modalities. Furthermore, the process results in imbalanced classification, complex computations, and high memory usage, making it less ideal for fusing spatiotemporal features and datasets.

Table 1: Summary of audio-video based traditional fusion and other fusion's related work.

Fusion	Feature extraction model	Modality	Datasets	Remarks
Early fusion [43]	LSTM	Audio-video	CMU-MOSEI	Sensitive to noise and misalignment between audio and video signals
Late fusion [43]	LSTM	Audio-video	CMU-MOSEI	High computational cost; less effective in modeling complex interactions between modalities
Hybrid fusion [44]	VGG-net	Audio-video	IIT-R SIER	Increased model complexity; risk of overfitting with limited data
Multimodal Factorization Model (MFM) [43]	Bayesian network	Audio-video	CMU-MOSEI	Computationally expensive; less scalable for large datasets
Graph-MFN (G-MFN) [45]	LSTM	Audio-video	CMU-MOSEI	Limited scalability
Multiplicative fusion (M3ER) [46]	LSTM	Audio-video	IEMOCAP, CMU-MOSEI	Prone to overfitting
Cross-Attention fusion [39]	Attention & concatenation	Audio-video	RAVDESS	Requires large amounts of data for effective attention training; sensitive to missing modality information
Transformer fusion [42]	Transformer-based pre-trained model	Audio-video	MELD, IEMOCAP, CMU-MOSEI	High memory consumption; needs extensive pretraining and large datasets
Multimodal fusion [47]	CNN	Audio-video	AVEC2017	Limited ability to capture temporal relationships;
Model level fusion [48]	2-layer LSTM	Audio-video	RECOLA	fine-tuning requires careful parameter tuning.
Tensor fusion network TFN [49]	Three-fold Cartesian product	Audio-video	CMU-MOSEI	Tensor-based fusion can be computationally prohibitive; sensitive to missing or noisy data.
Multimodal Dynamic Fusion Network [50]	Bi-directional gated recurrent unit (BiGRU)	Audio-video	IEMOCAP, MELD	Complex training process; BiGRUs can suffer from vanishing gradient problems with long sequences.

2.3 Research gap

Problem: Through a comprehensive review of the literature, we have gained crucial insights into audio and video feature extraction, various traditional cross-modal feature fusions (such as early, late hybrid, attention, and transformer fusion), and deep learning models, along with their comparative performances on benchmark multimodal datasets. Traditional fusion faces challenges with high dimensionality in large datasets, fails to optimize the interdependencies of features, and struggles with fine-grained emotion classification. However, a critical research gap still needs to be addressed to improve further, specifically to reduce the computational error in traditional fusion mechanisms for audio-video based emotion recognition systems and enhance their sustainability.

Solution: To address this gap, we propose a gating unit, and advanced cross-modal fusion mechanism (factorized bilinear pooling and compact bilinear pooling) as an alternative to traditional methods. This approach employs 2D CNN and 3D CNN simple deep neural network architectures to avoid pruning and quantizing the mode while managing insignificant weights and neurons. This solution can optimize the computational efficiency while maintaining high performance, contributing to the development of more sustainable and scalable emotion recognition systems.

3 Material and methods

In this section, we first describe the cross-modal fusion mechanism and its architecture. Next, we introduce three advanced cross-modal fusion mechanisms and its algorithm to enhance audio-video based emotion recognition from audio and video modality. Finally, we discuss the comparative performance of these techniques against state-of-the-art fusion mechanisms.

3.1 Cross-modal fusion mechanism

Cross-modal fusion is an effective technique for emotion recognition that involves extracting meaningful and essential features from two or more heterogeneous input sources or modalities using feature extraction processes, integrating these features, and subsequently training a deep learning model. This technique has contributed to many applications including emotion recognition and has continually evolved, demonstrating its versatility and effectiveness. Notably, cross-modal fusion has been successfully applied in many applications such as object detection [51], night pedestrian detection [52] , low light image semantic segmentation [53], and depression detection [54]. Cross-modal fusion mechanism intends to develop a joint representation that gathers all the collective essential features from all the modalities and feeds into a single vector while retaining each modality's contributions.

While traditional cross-modal fusion mechanisms are discussed in the literature review section, this section focuses on three advanced cross-modal fusion mechanisms for emotion recognition: Factorized bilinear pooling (FBP), Compact bilinear pooling (CBP), and Compact Bilinear Gated Pooling (CBGP).

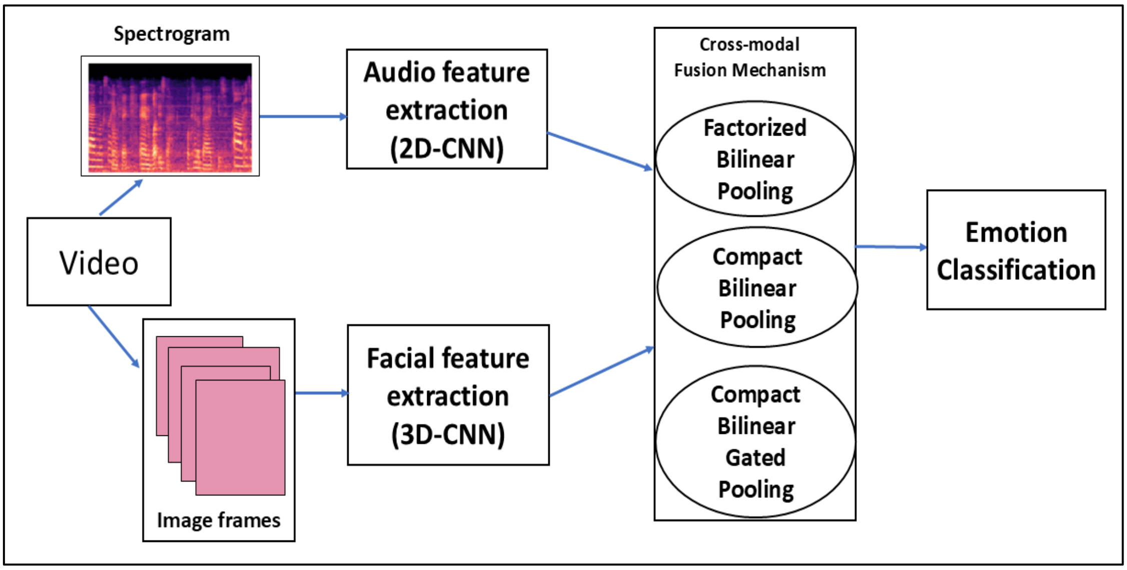
Spectrogram
Audio feature extraction (2D-CNN)
Video
Image frames
Facial feature extraction (3D-CNN)
Cross-modal Fusion Mechanism
Factorized Bilinear Pooling
Compact Bilinear Pooling
Compact Bilinear Gated Pooling
Emotion Classification

Figure 1: Basic architecture of Audio-video based cross-modal fusion mechanism

3.2 Factorized bilinear pooling (FBP)

Factorized Bilinear Pooling (FBP) is a method that enhances the standard bilinear pooling technique by factorizing the bilinear interaction tensor into lower-rank approximations [55]. Traditional bilinear pooling involves computing the outer product of two feature vectors from different modalities, resulting in a high-dimensional feature representation. While this method captures rich interactions between the modalities, it is computationally expensive and prone to overfitting due to the large number of parameters. FBP mitigates these issues by factorizing the interaction tensor into a product of two lower-rank matrices, significantly reducing the number of parameters while preserving the expressive power of bilinear interactions.

$$Z = \sum_{i=1}^{m} (M^T A).(N^T V) \qquad (1)$$

Where, Z: Pooled feature vector, M and N are bilinear interaction matrices, A and V are feature vectors from audio and video, respectively. Algorithm 1 illustrates the step-by-step factorized bilinear pooling fusion process implementation.

Training method: Let A' represents the Audio and V' represents the Video modality. The feature extraction functions $f_a \, and \, f_v$ are applied to the audio and video modality. It generates the feature vectors:

$$F_A = f_a(A') \; and \; F_V = f_v(V') \qquad (2)$$

Where F_A *and* F_V are the extracted feature vectors from audio and video, D_A *and* D_V are dimensionality spaces of the audio and video feature spaces.

If the features from audio and video need to be combined, a fusion mechanism Σ can be used to integrate these feature vectors into a unified representation F'. It can be calculated as:

$$F' = \Sigma(F_A, F_V) \;\; or \; F = F_A \oplus F_V \qquad (3)$$

Algorithm 1: Factorized Bilinear Pooling (FBP)

```
Input: Factorize audio features: $F_A = f_a(A')$
       Factorize video features: $F_V = f_v(V')$
Output: Predict the emotion class for new inputs
1. Compute the bilinear interaction between the
   factorized audio and video features:
   $F' = \Sigma(F_A, F_V) \;\; or \; F = F_A \oplus F_V$
2. Feed the compact bilinear pooled vector $Z_{FBP}$
   into a deep neural network classifier: $(F^i y^i)_{i=1}^N$
3. Calculate the loss function, minimize, and
   evaluation metrics
4. Use the trained model to predict the emotion
:lass
   for new inputs.
```

This factorization reduces the computational burden and allows the model to generalize better, especially when dealing with limited data. FBP has been successfully applied in tasks such as Visual Question Answering (VQA) and image-text matching, where the interaction between modalities is crucial.

3.3 Compact bilinear pooling (CBP)

Compact Bilinear Pooling (CBP) further refines the bilinear pooling approach by employing compact representations of the bilinear interactions. Unlike standard bilinear pooling, which directly computes the outer product of two feature vectors, CBP leverages approximations based on the Tensor Sketch technique to produce a compact representation of the outer product. This method dramatically reduces the dimensionality of the resulting feature vector without losing the key interactions between modalities. Algorithm 2 illustrates the compact bilinear pooling fusion process implementation.

In CBP, the outer product of the feature vectors A and V is approximated by projecting both vectors into a higher-dimensional space using random projections, followed by element-wise multiplication and summation. Presented equation represents how to implement CBP for audio-video emotion recognition using a deep neural network:

$$Z = \sum_{i=1}^{m}(proj_a(A)_i) \,.\, (proj_v(V_i)) \qquad (5)$$

A prediction function f (F') is then applied to the feature vector F' to predict the target emotion category value, Z' so, $Z' = f(F)$. Here, $f(F)$ is a 2D-CNN deep neural network acting as a classifier. The model is trained on labelled dataset so it is calculated as follows:

$(F^i y^i)_{i=1}^N$ (4) Where y^i is the true prediction label and N is the size of the sample.

Where, Z: Pooled feature vector, A and V are feature vectors from audio and video, respectively. $proj_a$, and $proj_v$ are projection matrix of audio and video features.

Training Method: Let A' represents the Audio and V' represents the Video modality. The feature extraction functions f_a *and* f_v are applied to the audio and video modality. It generates the feature vectors:

$$F_A = f_a(A') \; and \; F_V = f_v(V') \qquad (6)$$

CBP uses random projections to project the high-dimensional feature vectors into a lower-dimensional space before combining them. Random projection for audio (Z_A) and video features (Z_V):

$$Z_A = (P_A F_A) \text{ and } Z_V = (P_V F_V) \qquad (7)$$

Where, $Z_{A/V}$: Projection of audio /video, P_A, and P_v: Projection matrix of audio /video features. To maintain the information during projection, random sign vectors are applied to the projected features.

$$ZA' = SA \circ ZA \text{ and } ZV' = SV \circ ZV \qquad (8)$$

SA and SV are random sign vectors for audio and video features, $\circ$ *denotes element wise multiplication.* Then we applied the random permutation to the elements of the signed vectors to further scramble the features.

$$Z_{A''} = Permute(Z'_A, h_A), \; and \; Z_{V''} = Permute(Z'_V, h_V) \; (9)$$

where, h_A, h_V is a permutation vector applied to the indices of Z'_A and Z'_V.

The core of CBP involves computing the circular convolution of the two permuted feature vectors:

$$Z_{CBP} = FFT^{-1}\left(FFT(Z_A)\right) \circ \left(FFT(Z_V)\right) \qquad (10)$$

FFT^{-1} : *inverse fast fourier transform,* and

FFT: *fast fourier transform*

After this we normalized the obtained CBP feature vector and classified the categories of emotions by using deep neural networks. Calculated with the following formula:

$$Z'_{CBP} = softmax(Z_{CBP}) \quad (11)$$

where Z'_{CBP} predicts the emotion class, and $softmax(Z_{CBP})$ represents the function of the deep neural network.

```
Algorithm 2: Compact Bilinear Pooling (CBP)
Input: Project audio features, $Z_A$ $and$
       Project video features, $Z_V$
Output: Predict the emotion class for new inputs
1. Generate projection matrix, $ZA'$
2. Apply sign vectors to the projected audio features
       $ZA' = SA \circ ZA$
3. Apply sign vectors to the projected video features
       $ZV' = SV \circ ZV$
4. Apply permutation to the audio features:
       $Z_{A''} = Permute(Z'_A, h_A)$
5. Apply permutation to the video features:
       $Z_{V''} = Permute(Z'_V, h_V)$
6. Compute the circular convolution of the two permuted feature
   vectors: $Z_{CBP} = FFT^{-1}\left(FFT(Z_A)\right) \circ \left(FFT(Z_V)\right)$
7. Feed the compact bilinear pooled vector Z_CBP into a deep neural network classifier:
       $Z'_{CBP} = softmax(Z_{CBP})$
8. Calculate the loss function, minimize, and
       evaluation metrics
9. Use the trained model to predict the emotion class
   for new inputs.
```

3.4 Compact bilinear gated pooling (CBGP*)*

Compact Bilinear Gated Pooling (CBGP) enhances and builds upon Compact Bilinear Pooling (CBP) by adding a gating mechanism that adjusts or selectively emphasizes features based on their relevance, using a learned Softmax function to modulate feature interactions before pooling.

In CBGP, the feature vectors A and V undergo compact bilinear pooling as described in CBP, but before the final summation, the resulting interaction vector is element-wise multiplied by a gating vector $G' \in R^d$ Where, d is the dimensionality of the compact representation. The gating vector is computed as:

$$G' = \sigma(W_G(A', V') + b_G) \quad (12)$$

Where σ: $softmax\ function$, W_G : weight matrix, b_G: bias vector, A', V': audio, video feature vectors.

Training Method: Let A' represents the Audio and V' represents the Video modality. The feature extraction functions $f_a\ and\ f_v$ are applied to the audio and video modality. It generates the feature vectors:

$$F_A = f_a(A') \text{ and } F_v = f_v(V') \quad (13)$$

CBP uses random projections to project the high-dimensional feature vectors into a lower-dimensional space before combining them and calculates the random projection for audio (Z_A) and video features (Z_V):

$$Z_A = P_A F_A \ \ and \ \ Z_V = P_V F_V \quad (14)$$

Where Z_A $and\ Z_V$: Projection of audio and video, P_A, and P_V: Projection matrix of audio and video features. Then, we compute element-wise multiplication of the projected vectors:

$$Z' = Z_A \circ Z_V \quad (15)$$

Gated pooling: (i) compute the introduced gating vector $G' \in R^d$ Where, d is the dimensionality of the compact representation. The gating vector is computed as:

$$G' = \sigma(W_G(A', V') + b_G) \quad (16)$$

Here, σ is a Softmax function.

(ii) then we apply the gating mechanism to the element-wise multiply vector:

$$Z'' = G' \circ Z' \quad (17)$$

Finally, we sum the elements of the gated interaction vector to obtain the final pooled vector by the following equation-

$$Z = Sum(Z'') \quad (18)$$

This entire mechanism can be summarized by an equation, Where, Z: pooled feature vector, Z_A/$and\ Z_V$: Projection of audio and video.

$$Z = \sum_{i=1}^{m} \left(\sigma(W_G\ (A, V) + b_G)\right)_i . (Z_A)_i) . (Z_V(V_i)) \quad (19)$$

```
Algorithm 3: Compact Bilinear Gated Pooling (CBGP)
Input: Project audio features: Z_A, and Project video features: $Z_V$
Output: Predict the emotion class for new inputs.
1. Compute gating vectors for audio and video features:
       $G' = \sigma(W_G(A', V') + b_G)$
```

```
2.  Apply the gating vectors to the projected features:
        $Z' = Z_A \circ Z_V$
3.  Apply sign vectors to the gated audio features:
        $ZA' = SA \circ ZA$
4.  Apply sign vectors to the gated video features:
        $ZV' = SV \circ ZV$
5.  Apply permutation to the gated and signed audio features:
        $Z_{CBGP} = FFT^{-1}\ (P(FFT(Z_A \cdot G)))$
6.  Apply permutation to the gated and signed video features:
        $Z_{CBGP} = FFT^{-1}\ (P(FFT(Z_V \cdot G)))$
7.  Compute the circular convolution of the two permuted feature
    vectors:
        $Z'' = G' \circ Z'$
8.  Normalize the pooled feature vector:
        $Z = Sum(Z'')$
9.  Feed the compact bilinear gated pooled vector Z_CBGP into a deep
    neural network classifier:
        $Z_{CBGP} = \sum_{i=1}^{m} (\sigma(W_G\ (A, V) + b_G))_i . (Z_A)_i) . (Z_V(V_i))$
10. Calculate the loss function, minimize, and evaluation metrics
11. Use the trained model to predict the emotion class for new inputs
```

Through this mathematical analysis, CBGP has been able to identify the optimal fusion approaches that can be applied to audio-video-based emotion recognition systems, ultimately contributing to the development of more robust and accurate emotion recognition technologies. The gating mechanism allows to control the flow of information between the layers while selecting and rejecting the relevant or non-relevant (based on correlation feature score) inputs. As we know, not all the features are equally important at every step or time frame, so the gating mechanisms dynamically assign weights to features to capture complex regions more effectively.

4 Model training and experiments

Our experiments are conducted on a system equipped with an AMD Ryzen 7 processor, 16GB of RAM, and an NVIDIA GeForce RTX GPU. The code was implemented using Jupyter Notebook IDE and the PyTorch framework. For audio and video preprocessing, we utilized the Librosa and OpenCV Python libraries.

4.1 Evaluation dataset

CMU-MOSEI [37]**:** CMU-MOSEI dataset comprises over 23,259 annotated video clips collected from more than 1,000 speakers across a diverse range of topics. Total number of videos is 3228, video clips contain naturally occurring monologues in English, making the dataset a realistic representation of human communication. The dataset is annotated with six categorical emotions: happy, sad, angry, fear, disgusted, and surprised. Additionally, CMU-MOSEI provides intensity scores for each emotion, allowing for a fine-grained analysis of emotional expressions. After preprocessing, 20,323 samples are processed for feature extraction. The dataset is divided into three sets; 80% for training, 10% for testing, and 10% for validation. The performance is evaluated on Accuracy, F1-score, and mean absolute error, (MAE).

4.2 Deep learning model implementation details

a. 2D-CNN for Audio feature extraction and training model

We used 2D-cnn to extract and capture inter-modal feature dependencies from the CMU-MOSEI dataset. To generate spectrograms from raw audio files, we used the LibROSA library, which converts the raw audio waveform into a time series sampled at 22500 Hz. The waveform is then transformed into a spectrogram using the Short-Time Fourier Transform (STFT), with a window size of 2048 and a hop length of 512, striking a balance between time and frequency resolution. Spectrograms play a crucial role in audio-video emotion recognition as they align with video frames, increasing the likelihood of feature correlations due to time and frequency samples during fusion mechanism.

b. 3D-CNN for video feature extraction and training model

We used a simple 3D-CNN model because emotion recognition requires synchronized feature relations in each frame of a video, and a compact bilinear gated fusion mechanism can increase computational complexity. Additionally, our proposed approach aims to extract spatial and temporal features and incorporates a gated filter to fuse features from the audio and video modalities for each utterance. Therefore, we chose a simple deep learning architecture. The 3D-CNN takes a 224x224x3 image as input, which passes through the first 3D convolution layer followed by pooling layers, with a filter size of 3x3x3 and a stride of 1. Table 2 illustrates the Hyperparameters for 2D-CNN and 3D-CNN model.

Table 2. Hyperparameters for 2D-CNN and 3D-CNN model

Hyperparameter (2D-CNN) Audio	**Hyperparameter (3D-CNN) Video**
Input size= 224x224 Spectrogram	Input size=224x224x3 image frames
Kernels (conv layers) =32,64,128,256	Kernels (conv layers) = 64,128,256,512
Stride=1	Stride=1
Activation function= Relu and Softmax	Activation function= Relu and Softmax
Max Pooling= 2x2	Max Pooling= 3x3x3, 2x2x2
Batch size=32	Batch size=32
Epochs= 30	Epochs= 30-50
Learning rate=0.00003 (cosine decay)	Learning rate=0.00003 (cosine decay)
Regularization=L2	Regularization= L2
Dropout= 0.3%	Dropout=0.2%
Optimizer = Adam	Optimizer = Adam

5 Result and discussion

We evaluate the performance of each cross-modal fusion mechanism (FBP, CBP,CBGP) and compare it with the state-of-the-art (early fusion, late fusion and hybrid fusion) mechanisms on the CMU-MOSEI dataset using accuracy, F1-score, and MAE. F1-Score is the harmonic mean of precision and recall metrics. The results are summarized in the tables below, highlighting the contributions of each fusion method to the overall system performance

5.1 Ablation study

To investigate the specific contributions of compact bilinear gated pooling fusion (CBGP) of cross-modal fusion mechanism, this paper presents a detailed analysis of a series of ablation experiments conducted on the CMU-MOSEI datasets. These results are presented in tables, comparing key performance using accuracy, F1-score, and MAE among advanced cross-modal fusion mechanisms such as bilinear gated pooling, compact bilinear pooling, and compact bilinear gated pooling. We analyse the accuracy of each traditional fusion mechanism such as early fusion, late fusion and hybrid fusion on the same dataset, CMU-MOSEI. This approach employs 2D CNN and 3D CNN simple deep neural network architectures to avoid pruning and quantizing the mode while managing insignificant weights and neurons. The ablation study was carried out with a feature extraction process where features are audio and video modalities that interact through the outer product. The outer product allows the 2D-CNN and 3D-CNN to capture the interactions between every feature of one modality and every feature of the other modality in a compact manner. Comprehensive analysis and baseline comparisons show that our proposed CBGP fusion mechanism fuses feature effectively and outperforms the state-of-the-art fusion approaches. This study also provides a comprehensive discussion about transformer model based fusion approaches- attention fusion and transformer fusion.

5.2 Baseline comparisons

a. Comparison of advanced cross-modal fusion mechanism with state-of-the-art FBP, and CBP fusion mechanism.

Table 3: Performance comparison of advanced cross-modal fusion mechanisms on CMU- MOSEI dataset, highlighting their accuracy, F1-score, MAE, and specific strengths.

Cross-modal fusion mechanism	Accuracy (%)	F1-Score (%)	MAE	Remarks
FBP	76.9	75.6	59.1	Performs well with sentiment-emotion overlap
CBP	78.4	77.1	59.8	Captures diverse emotions effectively
CBGP	**80.3**	**79.2**	**54.2**	**Best for fine-grained emotion detection**

Table 3 illustrates that CBGP achieves the highest scores, particularly excelling in recognizing fine-grained emotions. Its ability to dynamically adjust the importance of different feature interactions allows it to handle the nuanced and varied expressions found in the illustrations in the CMU-MOSEI dataset.

b. Comparison of advanced cross-modal fusion mechanism with baseline cross-modal fusion mechanism

Table 4: Performance comparison of advanced cross-modal fusion mechanism with traditional, and baseline cross-modal fusion mechanism on CMU-MOSEI dataset, highlighting their accuracy, F1-score, and MAE.

Fusion Mechanism	Accuracy (%)	F1-score (%)	MAE (%)
Early fusion (EF-LSTM) [43]	78.2	77.9	64.2
Late fusion (LF-LSTM) [43]	80.6	80.6	61.9
Graph-MFN [45]	76.9	77.0	-
HFU-BERT model [56]	73.2	72.0	86.7
Early Fusion 2D-CNN (Ours)	67.3	65.4	69.7
Late Fusion 2D-CNN (Ours)	70.4	69.2	67.4
Hybrid Fusion 2D-CNN (Ours)	72.6	71.4	65.8
FBP (Ours)	76.9	75.6	59.1
CBP (Ours)	78.4	77.1	59.8
CBGP (Ours)	**81.3**	**79.2**	**54.2**

Table 4, illustrates that FBP performs well in scenarios involving sentiment-emotion overlap. CBP further improves by effectively capturing a diverse range of emotions. CBGP achieves the highest performance over traditional cross-modal fusion mechanisms due to limited feature interaction and correlation. CBGP excels in fine-grained emotion recognition and setting a benchmark on CMU-MOSEI dataset.

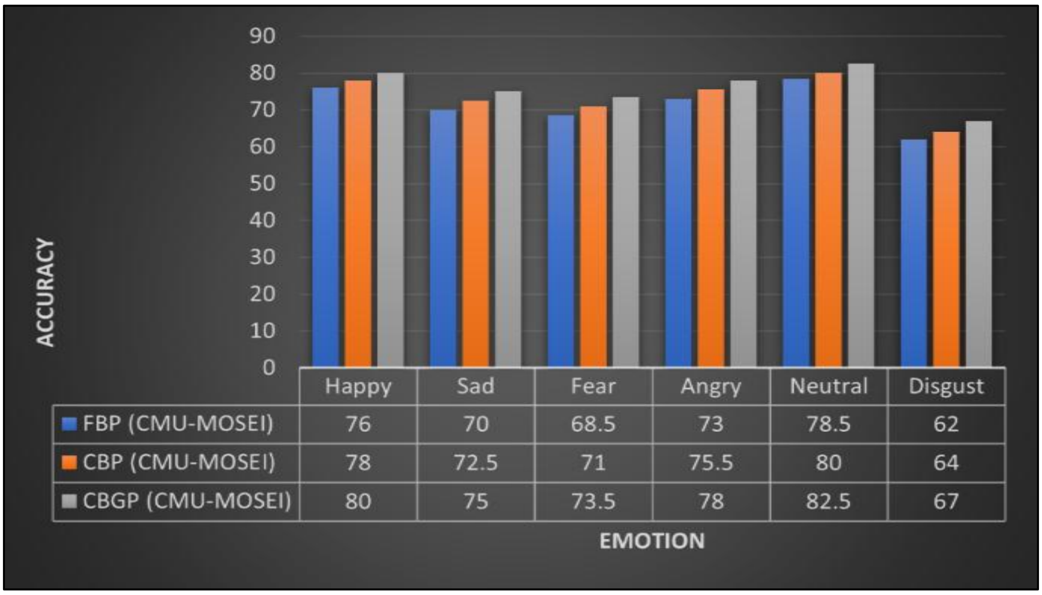

Figure 2: Accuracy performance of FBP, CBP, and CBGP fusion approaches on CMU-MOSEI

Figure 2 illustrates that in the CMU-MOSEI dataset emotion categories, CBP consistently outperforms FBP. The accuracy of 'Happy' emotion recognition increases from 76% (FBP) to 78% (CBP), and 'Sad' improves from 70% to 72.5%. CBGP provides higher accuracy than all other fusion mechanisms across all emotion categories. The progression from FBP to CBP, and from CBP to CBGP, emphasizes the strength and effectiveness of the fusion model in capturing emotional feature cues. This fusion leads to meaningful results that help classify emotion categories more accurately.

c. System complexity analysis

Table 5: Computational costs comparison (in floating point operations) for FBP, CBP, and CBGP approaches across CMU-MOSEI Datasets.

Datasets	FBP	CBP	CBGP
CMU-MOSEI	4.5×10^6	3.8×10^6	4.0×10^6

Table 5 presents the computational cost comparison, and highlights the relative efficiency of the FBP, CBP, and CBGP approaches on the CMU-MOSEI dataset. Despite the apparent efficiency of CBP, the marginal difference in computational costs, particularly the 0.2×10^6 FLOP gap between CBP and CBGP, raises questions about the trade-offs in performance. Lower computational costs may come at the expense of reduced accuracy or robustness in multimodal emotion recognition tasks. The slight increase in CBGP's computational load may reflect the additional overhead required to manage bi-modal interactions and graph-based modeling, potentially leading to enhanced performance and interpretability.

Table 6: Performance comparison of accuracy and p-value for cross-modal fusion mechanism.

Cross-modal fusion mechanism	Accuracy (%)	p-value
FBP	76.9	0.004
CBP	78.4	0.003
CBGP	**80.3**	**0.002**

Table 6 presents the accuracy and p-value of Full Bilinear Pooling (FBP), Compact Bilinear Pooling (CBP), and Compact Bilinear Gated Pooling (CBGP), where FBP achieves the lowest accuracy of 76.9%. CBP improves accuracy to 78.4% by introducing compact bilinear pooling. CBGP achieves the highest accuracy of 80.3% by incorporating the gating mechanism, which selectively emphasizes relevant features. The p-value decreases across the methods, indicating improved statistical significance with increasing accuracy. The values (0.004 for FBP, 0.003 for CBP, and 0.002 for CBGP) demonstrate that the performance improvements are statistically significant.

d. Comparison of CBGP fusion mechanism with attention fusion and transformer fusion

Transformer fusion: Transformer fusion is an advanced approach of fusion mechanism with the help of a pre-trained transformer model, which scales well to large datasets and long sequences due to parallel computations. This fusion is suitable for text-based emotion recognition tasks and natural language processing-based (NLP) applications because transformer fusion model such as BERT [57], RoBERTa [40] performs on all token embeddings parallelly which is not efficient to work with audio and video modalities together. Audio and video have large interdependencies of features and long sequences, as a result, the computational cost will be very high, training and testing will need more memory and computational burdens. Transformer fusion will also face challenges to extract, fuse and learn complex spatiotemporal features without architectural modifications in the model. Transformer fusion works by dividing the word sequences into tokens, which is feasible but if we divide long audio signals and high frame rate videos can lead to loss of important features, fine-grained temporal information, tokenization can reduce the effectiveness and increase the biases in SoftMax function.

Attention fusion: In our proposed work, we opted for CBGP over attention fusion to reduce the computational cost because the CMU-MOSEI dataset is largest dataset, and our proposed solution uses 2D-CNN for audio and 3D-CNN for video modalities to avoid pruning and quantizing the mode while managing insignificant weights and neurons. If we apply an attention fusion mechanism, we would need to apply self-attention fusion separately for both models and then integrate their outputs using multi-head attention fusion. This entire process would likely result in high dimensionality and an increased number of trainable parameters, leading to high memory usage and expensive computation.

Attention mechanism relies on element-wise scale dot products, which may cause high variance during training Since our implementation employs a simpler CNN architecture, in that case the model could predict unbalanced attention scores. The extreme parameters could further cause exponential computation issues, as unbalanced attention implies that the model may focus excessively on some regions while ignoring others. In conclusion, while attention fusion is an effective fusion mechanism, it is not a suitable fit for our employed deep learning emotion recognition model that’s why we have excluded it from the experiment. It may perform better with architectures such as fit well in ResNet [12], DenseNet [58], MobileNet [59], and other transformer-based models, where its capabilities can be better utilized.

5.3 Why CBGP outperforms better?

Representation capacity

Traditional fusion: Traditional fusion typically concatenate or aggregate features from multiple modalities, which can result in linear combinations of features, whereas attention and transformer fusion enhance inter-modality interactions by learning feature weights, but they still rely on additive or multiplicative relationships between modalities. They often struggle with complex feature interactions and fail to capture higher-order dependencies effectively.

Advanced fusion: Factorized bilinear and compact bilinear pooling can capture non-linear and higher-order interactions between features across modalities, which allows richer representations. These methods compress the high-dimensional feature space into a lower-dimensional representation while preserving inter-modal relationships, addressing the curse of dimensionality in traditional bilinear pooling.

Computational efficiency

Traditional fusion: Simple concatenation or weighted aggregation methods are computationally inexpensive but may lead to redundant or over-complex representations. Transformer-based fusion, although effective, can be computationally expensive due to quadratic complexity in multi-head attention over long sequences or large modalities.

Advanced fusion: Compact bilinear pooling and gated pooling introduce compact representations by leveraging approximations (e.g., Random Fourier Transform or Count Sketch). These methods significantly reduce computational and memory overhead compared to traditional bilinear pooling without losing important interaction features.

Dimensionality reduction

Traditional fusion: These methods often rely on post-fusion dimensionality reduction techniques (e.g., PCA) to manage high-dimensional outputs. However, these approaches are not integrated into the fusion process, potentially leading to loss of modality-specific information.

Advanced fusion: Methods like compact bilinear and gated pooling perform dimensionality reduction implicitly during fusion, ensuring that only the most relevant and informative interactions are preserved.

Modality-specific challenges

Traditional Fusion: Early and late fusion assume modalities contribute equally, potentially underperforming in scenarios where modalities have asymmetric importance or varying quality. Transformers address some modality-specific issues but may fail in noisy or sparse input scenarios without sufficient modality-specific pretraining.

Advanced fusion: Compact bilinear and gated pooling are robust to modality-specific variations. For example: Gated pooling introduces selective weighting mechanisms that dynamically prioritize certain modalities or features based on their relevance. Factorized pooling ensures that noisy or less-relevant features are naturally down-weighted during fusion.

Generalization and scalability

Traditional fusion: Simple techniques like early and late fusion can generalize well but may not scale effectively to high-dimensional, multimodal, or diverse datasets. Transformer-based fusion can scale better but may require large datasets and pretraining to perform effectively.

Advanced fusion: Advanced techniques like compact bilinear pooling generalize well to high-dimensional data and work effectively on smaller datasets due to efficient feature compression. Factorized approaches reduce overfitting by limiting parameter count, improving scalability to complex multi-modal systems.

Interpretability

Traditional fusion: Approaches like attention fusion or transformer-based fusion are somewhat interpretable due to explicit weighting schemes or attention maps. However, early and hybrid fusion methods lack interpretability since features are often combined in a black-box manner.

Advanced fusion: Compact bilinear pooling and gated pooling methods often lack explicit interpretability because the transformations (e.g., random projections, Fourier transforms) are more abstract.

Table 7: Comparison of FBP, CBP and CBGP based on various parameters.

Cross-Modal Fusion	Feature interaction level	Feature map dimensionality	Computation cost	Advantage	Limitation
FBP	Element-wise product	Reduced $k \ll d^2$	Low	Efficient approximation of bilinear pooling	Introduces small approximation errors.
CBP	Tensor sketching	Compact $k \ll d^2$	Medium	Balances efficiency and expressiveness	It does not capture the full bilinear interactions
CBGP	Selective Second order interaction	Compact $k \ll d^2$	Medium	Best for fine-grained classification, emphasizes key features	Require extensive hyperparameter tuning.

Table 7 discusses the performance of FBP, CBP and CBGP based on various parameters such as, feature interaction level, feature map dimensionality, computational cost, advantage and limitation various parameters. Here, d^2 represents the input feature dimensionality, and k is the dimensionality of the output representation in bilinear pooling. In CBP and CBGP, the value of k is important as it directly affects the trade-off between computational efficiency and model expressiveness. If k is lower than small memory needed but model may lose some effectiveness. Conversely, if k is higher, the model acts more expressively but the computational cost increases.

5.4 Real-time application

As we have seen in the above sections, CBGP has proven to be an effective fusion mechanism over traditional fusion mechanisms. This comprehensive study has demonstrated its full capability as cross-modal based emotion recognition. In real-time application, CBGP can extend beyond audio and video fusion. It can contribute significantly in audio-video-text based real-time applications as well. CBGP is a computationally effective and robust fusion mechanism, making it crucial to capture high correlation and relevant features for fusing heterogeneous modalities. Here are some real-time applications where CBGP can be applied in, Computer vision and pattern recognition, Natural language Process based language interactions, Recommendation systems for customer, Healthcare and medical applications, Robotics and automation system, Banking and E-commerce based digital applications, Security and surveillance based human safety application.

6 Conclusion & future scope

This study investigates the effectiveness of three advanced cross model fusion mechanisms; factorized bilinear pooling, compact bilinear pooling, and compact bilinear gated pooling for audio-video based emotion recognition. This comprehensive experiment is conducted on a widely recognized dataset; CMU-MOSEI. The gating mechanism integrated within CBGP enables the model to selectively emphasize relevant feature interactions, which is crucial for accurately recognizing complex and nuanced emotional expressions. We evaluated the performance of each fusion technique across various emotional categories, including happy, sad, fear, anger, neutral and disgust. The performance of advanced cross-modal fusion is compared to traditional cross-modal fusion mechanisms like early fusion, late fusion and hybrid fusion and transformer model based fusion mechanisms like, attention fusion and transformer fusion. The experimental results clearly demonstrate that the compact bilinear gated pooling (CBGP) mechanism outperforms the other fusion techniques across benchmark dataset, consistently achieving higher accuracy, F1-score, and MAE. Overall, the findings from this study suggest that incorporating a gating mechanism in multimodal fusion processes can significantly enhance the performance of emotion recognition systems, making CBGP a promising approach for future developments in this field.

References

[1] O. El Hammoumi, F. Benmarrakchi, N. Ouherrou, J. El Kafi, and A. El Hore, "Emotion Recognition in E-learning Systems," *6th Int. Conf. Multimed. Comput. Syst.*, pp. 1–6, 2018.

[2] Y. Zhang, Z. R. Wang, and J. Du, "Deep Fusion: An Attention Guided Factorized Bilinear Pooling for Audio-video Emotion Recognition," in *International Joint Conference on Neural Networks (IJCNN),* IEEE, 2019. doi: 10.1109/IJCNN.2019.8851942.

[3] Y. Li, X. Zheng, M. Zhu, J. Mei, Z. Chen, and Y. Tao, "Compact bilinear pooling and multi-loss network for social media multimodal classification," *Signal, Image Video Process.*, vol. 18, no. 11, pp. 8403–8412, 2024, doi: 10.1007/s11760-024-03482-w.

[4] D. Kiela, E. Grave, A. Joulin, and T. Mikolov, "Efficient large-scale multi-modal classification," *32nd AAAI Conf. Artif. Intell. AAAI 2018*, pp. 5198–5204, 2018, doi: 10.1609/aaai.v32i1.11945.

[5] W. A. Khan, H. ul Qudous, and A. A. Farhan, "Speech emotion recognition using feature fusion: a hybrid approach to deep learning," *Multimed. Tools Appl.*, vol. 83, no. 31, pp. 75557–75584, 2024, doi: 10.1007/s11042-024-18316-7.

[6] C. Yu, X. Zhao, Q. Zheng, P. Zhang, and X. You, "Hierarchical Bilinear Pooling for Fine-Grained Visual Recognition," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11220 LNCS, pp. 595–610, 2018, doi: 10.1007/978-3-030-01270-0_35.

[7] X. Peng, "Research on emotion recognition based on deep learning for mental health," *Inform.*, vol. 45, no. 1, pp. 127–132, 2021, doi: 10.31449/inf.v45i1.3424.

[8] B. Mocanu, R. Tapu, and T. Zaharia, "Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning," *Image Vis. Comput.*, vol. 133, p. 104676, 2023, doi: 10.1016/j.imavis.2023.104676.

[9] E. S. Salama, R. A. El-Khoribi, M. E. Shoman, and M. A. Wahby Shalaby, "A 3D-convolutional neural network framework with ensemble learning techniques for multi-modal emotion recognition," *Egypt. Informatics J.*, vol. 22, no. 2, pp. 167–176, 2021, doi: 10.1016/j.eij.2020.07.005.

[10] M. M. Hassan, M. G. R. Alam, M. Z. Uddin, S. Huda, A. Almogren, and G. Fortino, "Human emotion recognition using deep belief network architecture," *Inf. Fusion*, vol. 51, no. October 2018, pp. 10–18, 2019, doi: 10.1016/j.inffus.2018.10.009.

[11] L. Wang and J. Qiao, "Research and Application of Deep Belief Network Based on Local Binary Pattern and Improved Weight Initialization," in *3rd International Symposium on Autonomous Systems, ISAS 2019*, IEEE, 2019, pp. 1–6. doi: 10.1109/ISASS.2019.8757780.

[12] K. L. Lakshmi *et al.*, "Recognition of emotions in speech using deep CNN and RESNET," in *Soft Computing*, Springer Berlin Heidelberg, 2023. doi: 10.1007/s00500-023-07969-5.

[13] N. H. Ho, H. J. Yang, S. H. Kim, and G. Lee, "Multimodal Approach of Speech Emotion Recognition Using Multi-Level Multi-Head Fusion Attention-Based Recurrent Neural Network," in *IEEE Access*, IEEE, 2020, pp. 61672–61686. doi: 10.1109/ACCESS.2020.2984368.

[14] M. Sharafi, M. Yazdchi, R. Rasti, and F. Nasimi, "A novel spatio-temporal convolutional neural framework for multimodal emotion recognition," *Biomed. Signal Process. Control*, vol. 78, no. June, p. 103970, 2022, doi: 10.1016/j.bspc.2022.103970.

[15] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," *Appl. Acoust.*, vol. 158, p. 20, 2020, doi: 10.1016/j.apacoust.2019.107020.

[16] F. M. Alamgir and M. S. Alam, "Hybrid multi-modal emotion recognition framework based on InceptionV3DenseNet," *Multimed. Tools Appl.*, vol. 82, no. 26, pp. 40375–40402, 2023, doi: 10.1007/s11042-023-15066-w.

[17] S. K. Panda, A. K. Jena, M. R. Panda, and S.

Panda, "Speech emotion recognition using multimodal feature fusion with machine learning approach," *Multimed. Tools Appl.*, vol. 82, no. 27, pp. 42763–42781, 2023, doi: 10.1007/s11042-023-15275-3.

[18] S. W. Byun and S. P. Lee, "A study on a speech emotion recognition system with effective acoustic features using deep learning algorithms," *Appl. Sci.*, vol. 11, no. 4, pp. 1–15, 2021, doi: 10.3390/app11041890.

[19] H. Aouani and Y. Ben Ayed, "Speech Emotion Recognition with deep learning," in *24th International Conference on Knowledge-Based and Intelligent Information & Engineering Speech Emotion Recognition with deep learning Systems*, Elsevier B.V., 2020, pp. 251–260. doi: 10.1016/j.procs.2020.08.027.

[20] E. S. Agung, A. P. Rifai, and T. Wijayanto, "Image-based facial emotion recognition using convolutional neural network on emognition dataset," *Sci. Rep.*, vol. 14, no. 1, pp. 1–22, 2024, doi: 10.1038/s41598-024-65276-x.

[21] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, and A. Košir, "Audio-Visual Emotion Fusion(AVEF):A Deep Efficient Weighted Approach," *Inf. Fusion*, vol. 46, pp. 184–192, 2019, doi: 10.1016/j.inffus.2018.06.003.

[22] D. Ghimire, J. Lee, Z. N. Li, and S. Jeong, "Recognition of facial expressions based on salient geometric features and support vector machines," *Multimed. Tools Appl.*, vol. 76, no. 6, pp. 7921–7946, 2017, doi: 10.1007/s11042-016-3428-9.

[23] X. Yan, "A Face Recognition Method for Sports Video Based on Feature Fusion and Residual Recurrent Neural Network," *Inform.*, vol. 48, no. 12, pp. 137–152, 2024, doi: 10.31449/inf.v48i12.5968.

[24] S. R. Sanku and B. Sandhya, "Multi-Modal Emotion Recognition Feature Extraction and Data Fusion Methods Evaluation," *Int. J. Innov. Technol. Explor. Eng.*, vol. 3075, no. 10, pp. 18–27, 2024, doi: 10.35940/ijitee.J9968.13100924.

[25] T. Baltrusaitis, C. Ahuja, and L. P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, 2019, doi: 10.1109/TPAMI.2018.2798607.

[26] E. Ivanova and G. Borzunov, "Optimization of machine learning algorithm of emotion recognition in terms of human facial expressions," *Procedia Comput. Sci.*, vol. 169, no. 2019, pp. 244–248, 2020, doi: 10.1016/j.procs.2020.02.143.

[27] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Inf. Fusion*, vol. 59, pp. 103–126, Jul. 2020, doi: 10.1016/j.inffus.2020.01.011.

[28] S. Cunningham, H. Ridley, J. Weinel, and R. Picking, "Supervised machine learning for audio emotion recognition: Enhancing film sound design using audio features, regression models and artificial neural networks," *Pers. Ubiquitous Comput.*, vol. 25, no. 4, pp. 637–650, 2021, doi: 10.1007/s00779-020-01389-0.

[29] V. K. Sharma, "Designing of face recognition system," *Int. Conf. Intell. Comput. Control Syst. ICICCS 2019*, pp. 459–461, 2019, doi: 10.1109/ICCS45141.2019.9065373.

[30] S. Sahoo and A. Routray, "Emotion recognition from audio-visual data using rule based decision level fusion," *2016 IEEE Students' Technol. Symp. TechSym 2016*, pp. 7–12, 2017, doi: 10.1109/TechSym.2016.7872646.

[31] J. K. J. Julina and T. S. Sharmila, "Facial Emotion Recognition in Videos using HOG and LBP," in *2019 4th IEEE International Conference on Recent Trends on Electronics, Information, Communication and Technology, RTEICT 2019 - Proceedings*, IEEE, 2019, pp. 56–60. doi: 10.1109/RTEICT46194.2019.9016766.

[32] A. Vinay, V. S. Shekhar, K. N. B. Murthy, and S. Natarajan, "Face Recognition Using Gabor Wavelet Features with PCA and KPCA - A Comparative Study," in *3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015)*, Elsevier Masson SAS, 2015, pp. 650–659. doi: 10.1016/j.procs.2015.07.434.

[33] S. Kakuba, A. Poulose, and D. S. Han, "Deep Learning Approaches for Bimodal Speech Emotion Recognition: Advancements, Challenges, and a Multi-Learning Model," *IEEE Access*, vol. 11, pp. 113769–113789, 2023, doi: 10.1109/ACCESS.2023.3325037.

[34] X. Lu, "Deep Learning Based Emotion Recognition and Visualization of Figural Representation," *Front. Psychol.*, vol. 12, no. January, pp. 1–12, 2022, doi: 10.3389/fpsyg.2021.818833.

[35] M. Zielonka, A. Piastowski, A. Czyżewski, P. Nadachowski, M. Operlejn, and K. Kaczor, "Recognition of Emotions in Speech Using Convolutional Neural Networks on Different Datasets," *Electron.*, vol. 11, no. 22, 2022, doi: 10.3390/electronics11223831.

[36] K. Zhang, Y. Li, J. Wang, Z. Wang, and X. Li, "Feature fusion for multimodal emotion recognition based on deep canonical correlation analysis," *IEEE Signal Process. Lett.*, vol. 28, no. September 2022, pp. 1898–1902, 2021, doi: 10.1109/LSP.2021.3112314.

[37] C. Dixit and S. M. Satapathy, "Deep CNN with late fusion for real time multimodal emotion recognition," *Expert Syst. Appl.*, vol. 240, no. November 2023, p. 122579, 2024, doi: 10.1016/j.eswa.2023.122579.

[38] Y. Cimtay, E. Ekmekcioglu, and S. Caglar-Ozhan, "Cross-subject multimodal emotion recognition based on hybrid fusion," *IEEE*

Access, vol. 8, pp. 168865–168878, 2020, doi: 10.1109/ACCESS.2020.3023871.

[39] R. G. Praveen, E. Granger, and P. Cardinal, "Cross Attentional Audio-Visual Fusion for Dimensional Emotion Recognition," *Proc. - 2021 16th IEEE Int. Conf. Autom. Face Gesture Recognition, FG 2021*, 2021, doi: 10.1109/FG52635.2021.9667055.

[40] D. Sharma, M. Jayabalan, N. Sultanova, J. Mustafina, and D. N. L. Yao, "Multimodal Emotion Recognition Using Attention-Based Model with Language, Audio, and Video Modalities," *Lect. Notes Data Eng. Commun. Technol.*, vol. 191, pp. 193–210, 2024, doi: 10.1007/978-981-97-0293-0_15.

[41] Z. Fu *et al.*, "A cross-modal fusion network based on self-attention and residual structure for multimodal emotion recognition," pp. 2–6, 2021, [Online]. Available: http://arxiv.org/abs/2111.02172

[42] V. John and Y. Kawanishi, "Audio and Video-based Emotion Recognition using Multimodal Transformers," *Proc. - Int. Conf. Pattern Recognit.*, vol. 2022-Augus, no. August, pp. 2582–2588, 2022, doi: 10.1109/ICPR56361.2022.9956730.

[43] Y. H. H. Tsai, P. P. Liang, A. Zadeh, L. P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," *7th Int. Conf. Learn. Represent. ICLR 2019*, 2019.

[44] P. Kumar, S. Malik, and B. Raman, "Interpretable multimodal emotion recognition using hybrid fusion of speech and image data," *Multimed. Tools Appl.*, vol. 83, no. 10, pp. 28373–28394, 2024, doi: 10.1007/s11042-023-16443-1.

[45] P. P. Liang and R. Salakhutdinov, "Computational Modeling of Human Multimodal Language : The MOSEI Dataset and Interpretable Dynamic Fusion," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018. doi: 10.18653/v1/P18-1208.

[46] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3ER: Multiplicative Multimodal Emotion Recognition using Facial, Textual, and Speech Cues," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 1359–1367. doi: 10.1609/aaai.v34i02.5492.

[47] N. Singh, N. Singh, and A. Dhall, "Continuous Multimodal Emotion Recognition Approach for AVEC 2017," *Comput. Vis. Pattern Recognit.*, 2017, doi: 10.48550/arXiv.1709.05861.

[48] L. Schoneveld, A. Othmani, and H. Abdelkawy, "Leveraging recent advances in deep learning for audio-Visual emotion recognition," *Pattern Recognit. Lett.*, vol. 146, pp. 1–7, 2021, doi: 10.1016/j.patrec.2021.03.007.

[49] A. Zadeh, M. Chen, E. Cambria, S. Poria, and L. P. Morency, "Tensor fusion network for multimodal sentiment analysis," *EMNLP 2017 - Conf. Empir. Methods Nat. Lang. Process. Proc.*, pp. 1103–1114, 2017, doi: 10.18653/v1/d17-1115.

[50] D. Hu, X. Hou, L. Wei, L. Jiang, and Y. Mo, "Mm-Dfn: Multimodal Dynamic Fusion Network for Emotion Recognition in Conversations," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2022-May, pp. 7037–7041, 2022, doi: 10.1109/ICASSP43922.2022.9747397.

[51] A. R. Pathak, M. Pandey, and S. Rautaray, "Application of Deep Learning for Object Detection," *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 1706–1717, 2018, doi: 10.1016/j.procs.2018.05.144.

[52] Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian detection aided by deep learning semantic tasks," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 07-12-June, pp. 5079–5087, 2015, doi: 10.1109/CVPR.2015.7299143.

[53] A. H. Abdulwahhab, N. T. Mahmood, A. A. Mohammed, I. Myderrizi, and M. H. Al-Jumaili, "A Review on Medical Image Applications Based on Deep Learning Techniques," *J. Image Graph. Kingdom)*, vol. 12, no. 3, pp. 215–227, 2024, doi: 10.18178/JOIG.12.3.215-227.

[54] V. Adarsh, P. Arun Kumar, V. Lavanya, and G. R. Gangadharan, "Fair and Explainable Depression Detection in Social Media," *Inf. Process. Manag.*, vol. 60, no. 1, p. 103168, 2023, doi: 10.1016/j.ipm.2022.103168.

[55] H. Zhou, J. Du, Y. Zhang, Q. Wang, Q. F. Liu, and C. H. Lee, "Information Fusion in Attention Networks Using Adaptive and Multi-Level Factorized Bilinear Pooling for Audio-Visual Emotion Recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 2617–2629, 2021, doi: 10.1109/TASLP.2021.3096037.

[56] S. Lee, D. K. Han, and H. Ko, "Multimodal Emotion Recognition Fusion Analysis Adapting BERT with Heterogeneous Feature Unification," *IEEE Access*, vol. 9, pp. 94557–94572, 2021, doi: 10.1109/ACCESS.2021.3092735.

[57] S. Siriwardhana, T. Kaluarachchi, M. Billinghurst, and S. Nanayakkara, "Multimodal emotion recognition with transformer-based self supervised feature fusion," *IEEE Access*, vol. 8, pp. 176274–176285, 2020, doi: 10.1109/ACCESS.2020.3026823.

[58] M. A. H. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, and T. Shimamura, "Facial emotion recognition using transfer learning in the deep CNN," *Electron.*, vol. 10, no. 9, 2021, doi: 10.3390/electronics10091036.

[59] N. A. S. Badrulhisham and N. N. A. Mangshor, "Emotion Recognition Using Convolutional Neural Network (CNN)," *J. Phys. Conf. Ser.*,

vol. 1962, no. 1, 2021, doi: 10.1088/1742-6596/1962/1/012040.