




# Multimodal emotion recognition based on a fusion of audiovisual information with temporal dynamics

José Salas-Cáceres<sup>1</sup>  · Javier Lorenzo-Navarro<sup>1</sup> · David Freire-Obregón<sup>1</sup> · Modesto Castrillón-Santana<sup>1</sup>

Received: 20 December 2023 / Revised: 8 July 2024 / Accepted: 4 September 2024  
© The Author(s) 2024

## Abstract

In the Human-Machine Interactions (HMI) landscape, understanding user emotions is pivotal for elevating user experiences. This paper explores Facial Expression Recognition (FER) within HMI, employing a distinctive multimodal approach that integrates visual and auditory information. Recognizing the dynamic nature of HMI, where situations evolve, this study emphasizes continuous emotion analysis. This work assesses various fusion strategies that involve the addition to the main network of different architectures, such as autoencoders (AE) or an Embrace module, to combine the information of multiple biometric cues. In addition to the multimodal approach, this paper introduces a new architecture that prioritizes temporal dynamics by incorporating Long Short-Term Memory (LSTM) networks. The final proposal, which integrates different multimodal approaches with the temporal focus capabilities of the LSTM architecture, was tested across three public datasets: RAVDESS, SAVEE, and CREMA-D. It showcased state-of-the-art accuracy of 88.11%, 86.75%, and 80.27%, respectively, and outperformed other existing approaches.

**Keywords** Emotion recognition · Biometrics · Multimodal data fusion · Human-machine interaction

## 1 Introduction

The significance of emotions in human beings is widely acknowledged, as they exert influence over numerous facets of human behavior, decision-making, and even the perception of

---

✉ José Salas-Cáceres  
jose.salas@ulpgc.es

Javier Lorenzo-Navarro  
javier.lorenzo@ulpgc.es

David Freire-Obregón  
david.freire@ulpgc.es

Modesto Castrillón-Santana  
modesto.castrillon@ulpgc.es

<sup>1</sup> Universidad de Las Palmas de Gran Canaria, Instituto Universitario SIANI, Las Palmas de G.C., Spain

the world around us [25]. In a typical social interaction among humans, the emotional state of the individuals involved assumes considerable importance, as it delineates the conversational tone, the topics under discussion, and various other aspects thereof. In such situations, individuals exhibit a natural aptitude for detecting these emotional cues and adeptly adjusting their behavior in response. In [36], the author argues that all emotions can be characterized by two fundamental factors: valence, which relates to the degree of pleasure or displeasure, and arousal, which varies from low to high and denotes the intensity of physiological activation. Meanwhile, in [13] the authors emphasize that emotions rapidly prepare an individual for significant interpersonal encounters, and further notes that part of emotional behavior is innate, stemming from evolution, while part is acquired through learning. Ekman et al. also introduces criteria for identifying "basic" emotions, listing six of them in [12]: happiness, sadness, anger, fear, surprise, and disgust. Both works underscore the importance of recognizing emotions to anticipate an individual's behavior, and they acknowledge that cultural differences play a role in how people express emotions.

Additionally, in an era characterized by the growing presence of robots, human-machine interactions (HMI) are on the rise. Within this context, the quality of user experiences in HMIs is paramount for seamlessly integrating these technologies into society. Considering this, it becomes evident that the enhancement of HMI holds substantial significance. To achieve this improvement and fulfill a more natural interaction, it is imperative to endow machines with the capability to detect the user's emotions and to comprehend them, as described in [32]. Not all machines require emotional awareness, as some, such as computers, function effectively as rigid entities. Nevertheless, this capability can significantly enhance the user experience in numerous instances, allowing machines to adapt to the context, as argued in [33], with HMI as a pertinent example.

This work focuses on emotion detection, particularly facial expression recognition (FER). Similar to how humans perceive these behaviors through their visual and auditory senses in real-time, this research adopts a distinctively multimodal approach, integrating information from images and audio extracted from videos. The reason behind this dynamic method is rooted in the continuous nature of HMIs, where the state of the situation can, and often will, vary over time. Building on the observation that emotion recognition can be learned or given, and considering its significance in HMIs, this work aims to "teach" machines to perform this task. In other words, the objective is to design a method that can be integrated into robots, enabling them to recognize emotions and function in a future where HMIs are an important part of society. Additionally, this work aims to empirically verify the hypothesis that multimodal approaches often lead to better performance in these tasks compared to unimodal ones

The main contributions of this paper are as follows: a) the innovative integration of audio and video modalities within a novel architecture, b) the evaluation of various modality fusion strategies, and c) the execution of experiments across three public datasets, achieving state-of-the-art performance.

The structure of this paper is organized as follows: Section 2 discusses related work, Section 3 provides details about the databases employed in the experiments, Section 4 describes the proposed methodology and outlines the different experiments conducted, and finally, Section 5 presents and discusses the obtained results, with Section 6 offering concluding remarks.

## 2 Related work

Recently, within the field of FER, as in many other domains of Computer Vision, traditional methods such as Support Vector Machines (SVM) and logistic regression have yielded ground to new approaches founded on deep learning and neural network architectures as shown in [24, 39]. These contemporary techniques can be categorized in various ways. However, in this work, we will group them based on two key parameters: the amount of time the model can perceive and the number of modalities from which the features are extracted. The selection of these categorization criteria is motivated by the amount of temporal information processed (single image vs. sequences) and the number of modalities employed (unimodal vs. multimodal), which are fundamental aspects that significantly impact the design of these techniques. Furthermore, these factors directly influence the complexity of the methods applied and the type of information available for emotion recognition.

Regarding the temporal context in feature representation, these methods can be categorized into two classes as discussed in [24]: FER networks for static samples and FER networks for dynamic samples. The former focuses on processing data at a particular instant, such as static facial images, often incorporating complementary information like gender, age, or head poses from individuals to improve the results, as seen in [50]. In contrast, the dynamic approaches deal with extracting temporal information, utilizing methods like the interval temporal restricted Boltzmann machine in [47] or a 3D convolutional neural network in [23]. Other specialized elements for extracting temporal information include architectures like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), purpose-built for this task and easily integrable into various network types, as demonstrated in [44] and [46]. For both, extensions called Bidirectional LSTM (BiLSTM) and Bidirectional GRU (BiGRU) exist, which train two instances of either LSTMs or GRUs, one to process the sequence in the original direction and another in reverse, an example of their use can be found in [2], where they combine BiLSTM with other techniques such as convolutional networks to detect the pain intensity from facial expressions images. Another example of LSTMs in FER can be seen in [29], where the authors employ two LSTMs, one for audio and one for image, in order to extract features from videos and perform FER. Both categories share fundamental techniques, with a standard recommendation to use transfer learning and fine-tuning to address overfitting from limited databases. More recently, [16] proposed a robust FER classifier that performs well in challenging real-world environments by combining state-of-the-art models and a temporal-sequence classifier. Their classifier comprises three sequentially connected blocks. Each block is a sequential pipeline of a 1D convolutional layer, a batch normalization layer, and a ReLU activation.

The second classification criterion, which is based on the number of cues used, is suitable for virtually any biometric data application, and divides the methods into two categories: unimodal or multimodal. Unimodal networks rely on a single source of information and often heavily emphasize visual data, as demonstrated in architectures like the one described in [37]. Even when considering audio-related cues, unimodal approaches may involve creating a visual representation of the audio, such as a spectrogram, to extract information, as seen in [22]. However, this transformation does not alter the fact that the modality remains audio; since the initial data captured by the sensor is audio, this modality remains despite this processing step. In contrast, multimodal networks leverage various information fusion methods which, as highlighted in [42], can be performed at different levels. One approach is sensor-level fusion, where the data is combined at the time of acquisition. Alternatively, feature-level fusion, as in [14], is where a combination of modalities is produced at the network once the features have already been extracted. Beyond these, another way to fusion

multiple cues are score-level, as exemplified in [1], where the authors combine the scores of several models that learn from different information to perform a final decision. Finally, decision-level fusion, where different system methods are used to achieve multiple responses to the given task and then a voting-like system is used to select a single response, as the one seen in [26].

A significant challenge in FER research is the lack of datasets that capture real-world scenarios involving regular people. As noted in [30], existing datasets fall into two categories: laboratory-controlled datasets such as IEMOCAP by [4] or RECOLA by [35] and "in the wild" datasets like MELD by [34] or MUStARD by [5]. While controlled datasets provide valuable data, they may not fully represent the complexities of real-world scenarios due to the controlled environments. On the other hand, "in-the-wild" datasets, often extracted from TV shows or movies featuring professional actors, make the datasets not entirely in the wild, as the acquisition conditions are certainly controlled, therefore producing a lack of generalizability. These problems, as presented in [30], limit the performance when applying the models in real scenarios. While other classification schemes for FER datasets exist, the one adopted best exemplifies the previously noted concern.

As will be elucidated, the work presented here represents an application rooted in deep learning for processing dynamic multimodal data, using recommended techniques such as transfer learning and LSTMs. In this work, the multimodal fusion will be performed at a feature level using audio and visual information extracted from videos.

### 3 Datasets

This section provides an in-depth overview of the datasets employed in this work, offering detailed insights into the sources, composition, and characteristics of the data sources essential for our research. In addition, a table containing the different sets of emotions and other general information of each dataset can be seen in Table 1.

#### 3.1 RADVESS

The initial dataset under consideration is the RADVESS database, initially described in [27]. This dataset contains audio and video recordings of 24 professional actors conveying eight distinct emotions: neutral, calm, happy, sad, angry, fearful, disgust, and surprise, resulting in a total of 7356 recordings, where each actor perform 104 unique vocalizations. Each actor performed two recordings for each emotion, except for the "Neutral" emotion, for which only one recording was made. The dataset exhibits meticulous balance, with an equal distribution

**Table 1** Emotions considered in each databases

Database	# emotions	Resolution	Cues	# of persons
RAVDESS	8, (anger, calm, disgust, fearful, happiness, neutral, sadness, surprise)	1920x1080	Audio and voice	24
SAVEE	7, (anger, disgust, fear, happiness, neutral, sadness, surprise)	960x760	Audio and voice	4
CREMA-D	6, (anger, disgust, fear, happiness, neutral, sadness)	480x360	Audio and voice	91

of 12 female and 12 male actors, although it predominates caucasian individuals over other ethnic groups. Notably, this dataset is partitioned into two subsets: one containing samples of the actors singing and another involving actors speaking. Only the latter subset was utilized for this research, as it better emulates a realistic environment. RAVDESS provides two levels of intensity for each emotion, except neutral, normal, and strong, both used indistinctly as the same emotion. The videos were recorded in a professional recording studio with controlled light, high-quality cameras, and microphones. The videos range in duration from 3 to 5.5 seconds, exemplary frames can be observed in Fig. 1. RAVDESS's extensive use in research has established it as a reference dataset within the community [28]

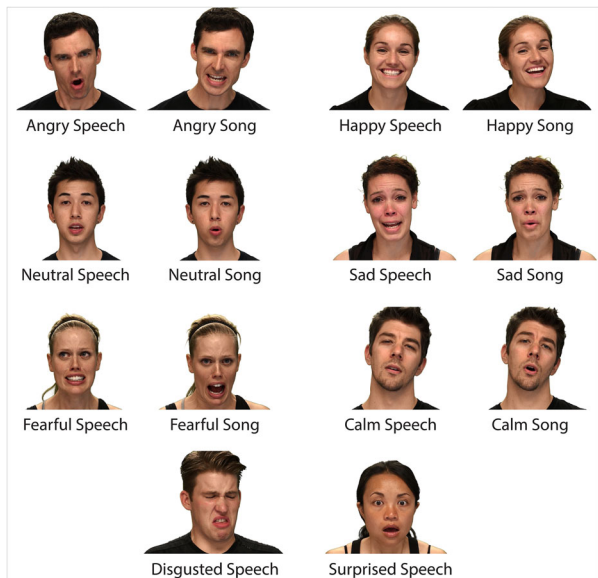
### 3.2 SAVEE

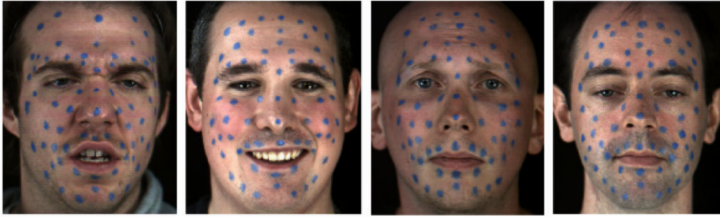
The second database employed in this study is the Surrey Audio-Visual Expressed Emotion (SAVEE) database by [18]. This dataset comprises recordings from four native English-speaking caucasian males aged between 27 and 31 years. As such, there is an inherent demographic bias within this dataset. Each individual has contributed 15 recordings for each of the seven distinct emotions, except for "neutral", which features 30 recordings. Among the 15 phrases recorded, three are consistent across all emotions, two are specific, and 10 are generic phrases, resulting in 120 recordings per individual. In order to extract facial expression features, the actor's faces were painted with 60 markers adequately distributed across the facial area. The film conditions were good, with proper illumination and a controlled environment. Figure 2 illustrates visual samples from the database.

### 3.3 CREMA-D

The CREMA-D database is a comprehensive repository containing 7,442 original clips delivered by 91 actors, 48 of which were males and 43 females as seen in [20]. Notably, this database exhibits deliberate balance in actor demographics, including gender, ethnic group, and age although as in the other described datasets, there is a prevalent appearance

**Fig. 1** Example of frames seen in RAVDESS. Extracted from [27]





**Fig. 2** Blue markers for tracking placed on subjects' faces with various emotions (from left): KL (angry), JK (happy), JE (sad) and DC (neutral) from the SAVEE database. Extracted from [17]

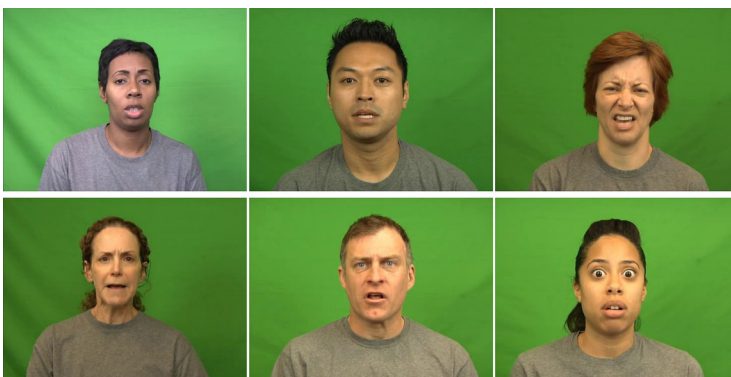
of caucasian actors. These actors articulated a set of 12 sentences, which were subsequently employed to elicit a wide range of emotional expressions, encompassing six emotions and four intensity levels. While the videos are divided into intensity levels, the experiment in this study exclusively focuses on the differentiation of emotions, utilizing the four intensity levels for each emotion as they were uniformly distributed. Even though the film conditions were controlled, the quality of the videos is compromised by the native 960x760 resolution of the camera used. Some CREMA-D frames can be seen in the Fig. 3.

## 4 Methodology

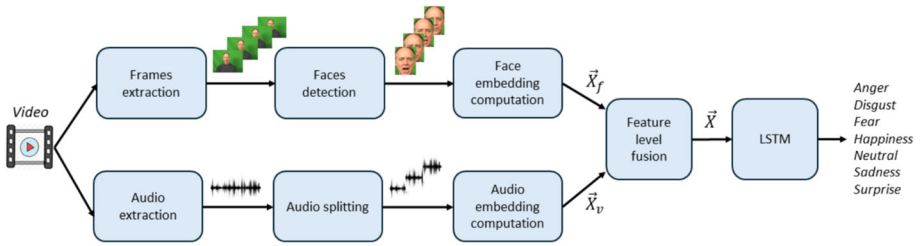
This section details the methodology employed in the experiments. We will delve into the data preparation process and the architectures used. Figure 4 visually illustrates the entire workflow. Section 4.1 will detail the data extraction and preparation steps, while Section 4.2 will present the architecture of the proposed models, including the considered multimodal fusion methods.

### 4.1 Data preparation

As previously explained, the proposed approach is inherently multimodal, emphasizing distinct biometric data sources for emotion recognition. The acquisition of these data commenced with a preprocessing step applied to the video clips, as visually depicted in Fig. 5. Given the presence of two primary information sources, namely audio and image, two preprocessing procedures were conducted, each starting with the original audiovisual format.



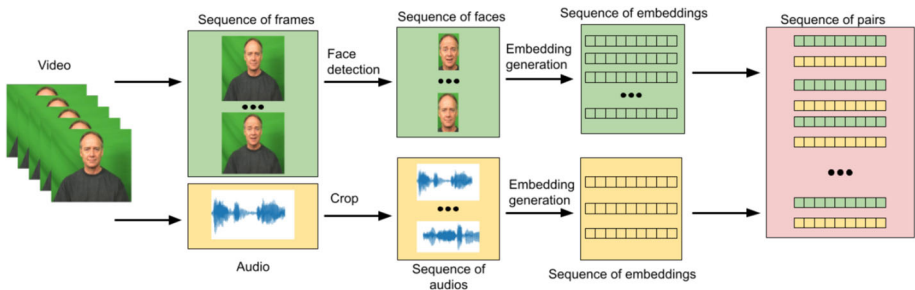
**Fig. 3** Example of frames seen in CREMA-D



**Fig. 4** Proposed architecture pipeline: frames and audio from the video are initially extracted. Subsequently, faces are detected within the frames, while the audio is segmented into chunks. Following this, embeddings are computed for each modality, which are then fused before being inputted into the LSTM classifier to recognize emotions

For the visual data, facial detection was performed on video frames using the Haar Cascade classifier [45]. The decision to use this model, instead of more recent ones like RetinaFace [10], LightFace [40], or MTCNN [49], is based on the controlled conditions of the videos, which consistently feature frontal close-ups of individuals’ faces, making face detection a simple task. From the detected faces, embeddings were generated using a facial embedder, in this case, VGG-Face with VGG-16 as its backbone architecture, see [31]. Notably, the VGG-Face embedder is highly regarded for its proven performance in numerous biometric tasks as the ones seen in [6], [15] or [48] and in numerous competitions. Then, two additional processes were applied: first, only the embeddings generated from the odd-numbered frames of the video were selected. This effectively halved the frame rate of the original videos, eliminating redundant information in the samples. Finally, to align with the LSTM architecture adopted, these vectors were grouped into sequences of 32, maintaining their chronological order of appearance in the original video. This process resulted in the creation of multiple sequences per video.

The procedure for preprocessing audio data is closely paralleled to that of the images. Initially, the audio was extracted from the video and segmented into three chunks to preserve temporal coherence. These fragments were subsequently aligned chronologically with the faces in the multimodal approaches. This alignment ensures that the first third of video frames corresponds to the first segment of audio, the second third to the second segment, and so on. Figure 6 shows a visual representation of this association. The audio embeddings were generated using an audio embedder, in this case, X-vector [43], explicitly trained to transform audio data of varying duration into fixed-dimensional vectors. Similar to the treatment of visual information, in the unimodal case, the generated vectors were structured in sequences.



**Fig. 5** Data preparation pipeline

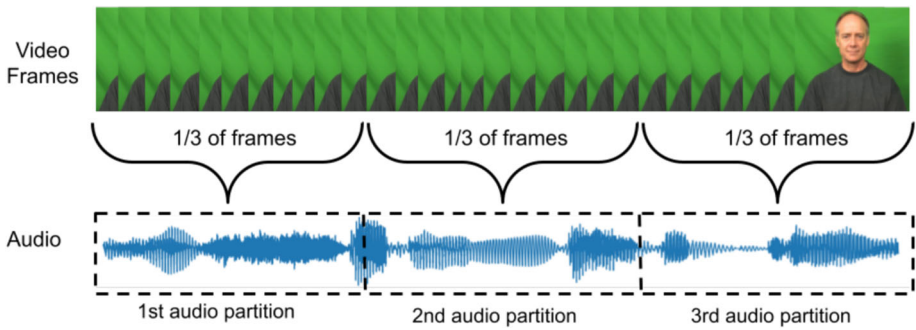


Fig. 6 Visual representation of the association video frame - audio segment

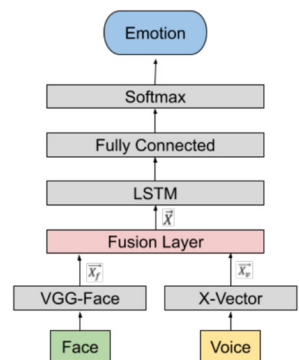
Due to the preprocessing steps described, visual and audio data were transformed into numerical vectors, making them more suitable for applying fusion techniques. The face embeddings possess dimensions of 2048, while the audio embeddings comprise 512 elements. These dimensions were defined by the size of the last layer of the embedding architecture.

The fusion of these two distinct data sources transpired at a feature level. This implies that the fusion process occurred within the feature vectors generated from both biometric sources, creating a multimodal application, as defined in [42].

### 4.2 Proposed architecture

Inspired by the architecture outlined by [3], the architecture proposed in this work is underpinned by two central concepts: a feature fusion layer that combines the cues and a LSTM layer capable of capturing dynamic features. A visual representation of this architecture is illustrated in Fig. 7. The input to the network consists of embeddings generated from the processes discussed in Section 4.1. Denoted as  $\vec{X}_f$  for face images and  $\vec{X}_v$  for audio, these inputs enter the fusion layer  $F$  and produce  $\vec{X} = F(\vec{X}_f, \vec{X}_v)$ , which corresponds to the input for the LSTM. The data then passes through a fully connected layer with its corresponding activation function (ReLU) and dropout, concluding with classification through a softmax layer.

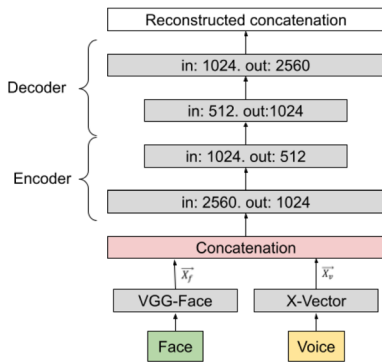
Fig. 7 Proposed architecture



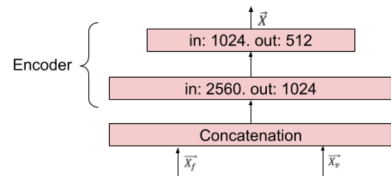


Four distinct fusion strategies were considered, each varying in how the feature fusion layer behaves:

- a) Visual approach: This approach exclusively utilizes video frames as the data source. So only the extracted information from the faces is fed up to the LSTM,  $F(\vec{X}_f, \vec{X}_v) = \vec{X}_f$ .
- b) Audio approach: Similarly to the previous approach, only audio sequences are employed; therefore,  $F(\vec{X}_f, \vec{X}_v) = \vec{X}_v$ . Both this strategy and the previous one can serve as a baseline for analyzing the performance of the remaining multimodal strategies.
- c) Multimodal concatenation: In this approach, the first true multimodal strategy, the audio feature vectors for each video are paired with the image embeddings, creating image-audio pairs. These pairs are formed by associating the first audio segment with the first third of the images, the second with the second third, and so on. Consequently, sets of 32 pairs are generated. Subsequently,  $F(\vec{X}_f, \vec{X}_v) = \vec{X}_f \oplus \vec{X}_v$ , where  $\oplus$  refer to the concatenation operator, converting each pair into a single vector with a length equivalent to the sum of their lengths.
- d) Multimodal Principal Component Analysis (PCA): The main problem of fusing by concatenation is the high dimensionality of the resulting vector  $\vec{X}$ . As shown in [21], PCA offers a solution to this problem, allowing the retention of essential information from both modalities without significant loss. It exploits the correlations between elements within the concatenated vector to extract the non-shared and most informative features. The PCA's hyperparameters were adjusted to ensure that the output vector retains 95% of the information present in the original. This resulted in dimensionality reductions to 423, 238, and 456 for the RAVDESS, SAVEE, and CREMA-D datasets, respectively, from their original dimensionality of 2560.
- e) Multimodal autoencoder: In this approach, the network builds upon the frame-audio sequences previously established. As stated and exemplified in [41], multimodal autoencoders (AEs) can be used as a way of extracting shared information between the cues and improving the fusion. Because of this, an AE is trained to reconstruct the pairs, with the AE's structure depicted in Fig. 8a. Once the AE is trained effectively, it is partitioned, with the encoder component isolated and incorporated into the network as  $F(\vec{X}_f, \vec{X}_v) = \text{encoder}(\vec{X}_f \oplus \vec{X}_v)$ , the structure is showed in Fig. 8b. This integration facilitates merging the two embeddings, resulting in a latent space encapsulating features from both modalities.



(a) Structure of the AE during training.



(b) Structure of the AE when working as a fusion layer.

**Fig. 8** AE structure in its training phase and when acting as a fusion layer

f) Embracement: The same data preparation steps are applied as in the previous multimodal approaches. However, in this case, what was previously an encoder or concatenation function is replaced by an EmbraceNet module [7]. EmbraceNet takes a set of input vectors and fuses them to generate a unified representation incorporating information from every cue. The model consists of a set of docking layers and a single embracement layer. Docking layers ensure that every input vector is the same size. This is achieved by applying a fully connected layer with the desired output size, followed by a ReLU activation function. In the embracement layer, these processed vectors are combined into a single, unified representation using a technique based on multinomial sampling. This technique ensures that the final vector remains the same size even if one of the input embeddings is missing. This module was trained concurrently with the rest of the network. Since during the training of the dockers layers, they consider the correlation between the different modalities, EmbraceNet's selection is motivated by the idea that exploiting these correlations leads to effective multimodal fusion.

The composition of  $\vec{X}$  for each architecture is detailed in Table 2. All architectures insert the visual embeddings first, followed by the audio embeddings. While the order of insertion affects the specific index assigned to each element within the final vector, it has no impact on the actual information content of it. This is because the methods employed exploit the correlations between elements from both modalities as a unit.

## 5 Experiments

This section details the application and evaluation of the methods described in Section 4. Section 5.1 outlines the training setup, while Section 5.2 explains the chosen evaluation metrics. Finally, Section 5.3 presents and discusses the obtained results. The conducted experiments considered the three datasets described in Section 3. Each dataset was divided into two subsets: a training set consisting of 80% of the samples and a test set comprising the remaining samples. The number of units in the LSTM layer was set to 16 and was trained for 100 epochs using the Adam optimizer. An adaptive learning rate was employed during training, with values adjusting between  $1e^{-03}$  and  $1e^{-05}$  based on the network's specific requirements. The number of outputs in the softmax layer corresponds to the number of emotions in the dataset.

Fusion strategies *Multimodal AE* and *Embracement*, described in Section 4.2, are based on trainable models, namely an AE and EmbraceNet. Therefore, they must be trained.

### 5.1 Training

In the training phase, the entire architecture depicted in Fig. 7 was trained, excluding the feature extractors, which were used with their original weights frozen. When using the Embracement strategy, the EmbraceNet network was trained alongside the other layers of the model. It was configured with an embracement size of 1024, indicating that it would convert the two embeddings into a single vector of that many elements.

In the particular case of the AE, the isolate architecture was trained independently with a concatenation of both cues to extract features and reconstruct the vector. The layers constituting the encoder were then extracted from the AE and integrated into the primary model as the fusion layer, with its parameters frozen. The training configuration used was the same as that employed in the previous approaches, using the same optimizer, hyperparameters,

**Table 2** Composition of  $\vec{X}$  in each approach

Approach	LSTM input ( $\vec{X}$ )
Visual	$\vec{X} = \vec{X}_f$
Aural	$\vec{X} = \vec{X}_v$
Concatenation	$\vec{X} = \vec{X}_f \oplus \vec{X}_v$
PCA	$\vec{X} = PCA(\vec{X}_f \oplus \vec{X}_v)$
AE	$\vec{X} = encoder(\vec{X}_f \oplus \vec{X}_v)$
Embracement	$\vec{X} = embracenet(\vec{X}_f \oplus \vec{X}_v)$

and data partitions. As stated before, the differences between the AE architecture during its training and when used as a fusion layer are visualized in Fig. 8.

## 5.2 Evaluation metrics

Two key evaluation metrics will be used to assess the performance of the proposed model comprehensively: Unweighted Average Recall (UAR) and Accuracy. UAR provides a balanced measure of the model's ability to correctly identify each emotion class, regardless of class distribution in the dataset [19, 37]. Accuracy, however, reflects the overall percentage of correctly classified samples. Utilizing both metrics ensures a well-rounded evaluation, considering both class-wise performance and overall classification accuracy.

## 5.3 Results and discussion

Table 3 shows the results obtained with unimodal and the proposed fusion strategies on the three datasets under consideration. In RADVESS and CREMA-D datasets, the *Embracement* strategy outperforms the other three. This is not the case for the SAVEE dataset, where the best result is obtained using only the visual modality, with the *PCA* strategy in second place. The other unimodal strategy, namely acoustic, yields the lowest accuracy in the three datasets. However, the *Concatenation* and *AE* multimodal strategies perform poorly. *PCA*'s performance across all three datasets aligns closely with the visual approach. This suggests that *PCA* might primarily extract the most informative features from the visual cues. This aligns with the observation that it achieves the second-best results in the SAVEE dataset, where visual information seems essential. In general, it can be seen that the multimodal approaches outperform or at least achieve similar results than the unimodal approaches, confirming the hypothesis that the multimodal approaches could lead to better performance than the unimodal ones.

The confusion matrices for the models with the best performance in each dataset are depicted in Fig. 9. Notably, in multimodal applications such as the ones used with RADVESS and CREMA-D, a prominent source of errors involves the confusion between sadness and fear. Furthermore, in RADVESS, there is an additional association between fear and anger, while in CREMA-D, this connection is seen between neutral and sadness. In contrast, the visual approach employed in SAVEE encounters two main issues: the confusion between happiness and surprise and the erroneous detection of neutral in sadness samples.

Table 4 compares the performance achieved with the proposed architecture and recent results from other literature studies. Notably, the proposed *Embracement* fusion strategy sets a new state-of-the-art benchmark in the RADVESS dataset, presenting competitive results in both CREMA-D and SAVEE datasets. While the result may not claim the best top result, it

**Table 3** Accuracy and UAR achieved by the used approaches in each database. Best results for each one of dataset are bolded

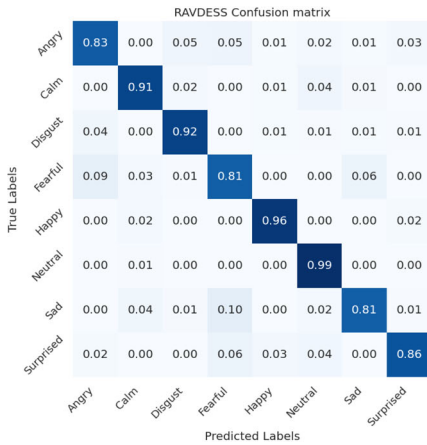
Approach	Accuracy (%)	UAR (%)
RAVDESS		
Visual	82.19	79.89
Acoustic	45.83	44.07
Multimodal concat	80.51	79.59
Multimodal PCA	82.68	83.38
Multimodal AE	73.66	73.70
Embracement	<b>88.11</b>	<b>88.56</b>
SAVEE		
Visual	<b>90.60</b>	<b>87.63</b>
Acoustic	60.42	50.78
Multimodal concat	77.62	74.91
Multimodal PCA	86.75	85.23
Multimodal AE	76.01	73.62
Embracement	82.13	79.77
CREMA-D		
Visual	73.97	74.05
Acoustic	42.01	42.50
Multimodal concat	71.34	70.72
Multimodal PCA	72.86	73.17
Multimodal AE	67.07	66.21
Embracement	<b>80.27</b>	<b>80.17</b>

closely rivals the performance of leading approaches. The reader must observe that for the unbalanced SAVEE dataset, an unimodal approach beats the *Embracement* fusion. In any case, the homogeneous behavior provided by *Embracement* fusion strategy evidences the proposed architecture's effectiveness, versatility, and competitive performance in the FER context.

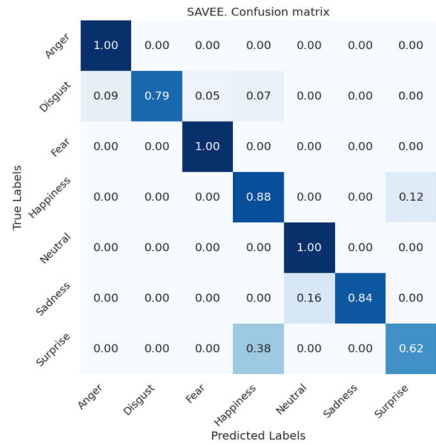
Our findings on the RAVDESS dataset are particularly remarkable. The strong performance of multimodal approaches on this dataset supports the hypothesis that these methods can outperform unimodal ones, at least when both modalities are of high quality, as with RAVDESS.

It is noteworthy that in the SAVEE dataset, the only dataset where the multimodal approach does not yield the best result, all the top scores are achieved by unimodal facial approaches. This observation aligns with the earlier assertion that the dataset's inherent bias makes it more reliant on visual information. This tendency may explain the superior performance of unimodal facial approaches in this particular dataset. This fact reveals the main limitation of the proposed multimodal approaches, where if one of the cues achieves significantly better performance than the others due to their qualities or quantity, the fusion could dilute the importance the network gives to the first one, resulting in a loss of performance. However, it is noteworthy that the multimodal approach achieves comparable results to these unimodal techniques.

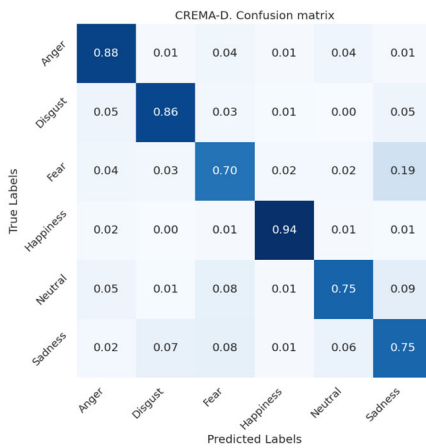
In contrast to the RAVDESS findings, the CREMA-D dataset reveals that unimodal approaches for both audio and visual modalities previously achieved the best results. Interest-



(a) RAVDESS.



(b) SAVEE.



(c) CREMA-D.

Fig. 9 Confusion matrices for the best model in each dataset

ingly, the multimodal EmbraceNet method achieves comparable performance on CREMA-D. These findings suggest that CREMA-D may be a suitable dataset for exploring multimodal approaches.

As perceived in the previous discussion, the proposed *EmbraceNet* fusion strategy achieve state-of-the-art results in each case. This success can be attributed to two key strengths, as stated in [7], the architecture inherently considerate the cross-modal correlation that exist between the cues when they come from the same source. Secondly, the architecture incorporates an internal dropout-like mechanism that effectively prevents overfitting to specific modalities during training. This mechanism distinguishes *EmbraceNet* from other fusion techniques. *PCA*, for instance, often prioritizes features from the dominant cue, not distinguishing itself from the best of the unimodal approaches, which can be observed in the difference between *PCA* results and those obtained using the unimodal visual cues from the RAVDESS and CREMA-D datasets as shown in Table 3. Meanwhile, the *Multimodal*

**Table 4** Comparative performances of state-of-the-art methods and the proposed architecture in each of the three datasets. Best results for each one of dataset are bolded

Ref. & year	Metric	Performance (%)	Modality
RAVDESS			
[8]	ACC	81.58	Multimodal (Audio + Image)
[38]	ACC	84.10	Unimodal (Audio)
[28]	ACC	86.70	Multimodal (Audio + Image)
Visual approach (Ours)	ACC	82.19	Unimodal (Image)
Multimodal embracement (Ours)	ACC	<b>88.10</b>	Multimodal (Audio + Image)
SAVEE			
[37]	UAR	82.8	Unimodal (Image)
[11]	ACC	86.5	Unimodal (Image)
Visual approach (Ours)	ACC	<b>90.6</b>	Unimodal (Image)
Multimodal embracement (Ours)	ACC	82.13	Multimodal (Audio + Image)
CREMA-D			
[9]	ACC	70.95	Unimodal (Audio)
[22]	ACC	<b>82.96</b>	Unimodal (Audio)
[37]	UAR	79.0	Unimodal (Image)
Visual approach (Ours)	ACC	73.97	Unimodal (Image)
Multimodal embracement (Ours)	ACC	80.27	Multimodal (Audio + Image)

ACC and UAR come from accuracy and unweighted average recall, respectively

*AE*, focuses on reconstructing the initial concatenation of both embeddings, not necessarily guaranteeing the extraction of task-relevant features. *Embracement*, on the other hand, effectively creates a fusion that often led to a better representation of the unimodal features and, therefore, better performance.

This mechanism distinguishes Embracement from other fusion techniques. PCA, for instance, often prioritizes features from the dominant modality, which can resemble a unimodal approach. Meanwhile, the Multimodal Autoencoder focuses on reconstructing the initial concatenation of both embeddings, not necessarily guaranteeing the extraction of task-relevant features. Embracement, on the other hand, effectively creates a fusion that consistently leads to superior performance.

All experiments were conducted using Python as the primary programming language. The training was performed on a computer with an Nvidia GTX 3080 10 GB graphics card, an Intel Core i7-11700K CPU (3.60 GHz), and Ubuntu 22.04 as the operating system. The architectures were created using the Pytorch library.

## 6 Conclusions

In conclusion, this study introduces a novel architecture for multimodal emotion recognition, comprising a fusion layer for integrating audio and visual cues at a feature level and an LSTM layer for capturing dynamic features. Notably, the architecture achieves state-of-the-art results in the RAVDESS and SAVEE datasets, underscoring its effectiveness in enhancing FER performance. Multimodal approaches, such as *Multimodal Concatenation* and *Multimodal AE*, consistently outperform unimodal approaches in datasets like RAVDESS

and CREMA-D, highlighting the benefits of leveraging the synergy between audio and visual cues. However, the study also emphasizes the influence of dataset characteristics, as observed in SAVEE, where a solid visual bias and audio quality limitations lead to unimodal facial approaches outperforming multimodal ones. This research underscores the versatility and adaptability of the proposed architecture across various datasets and emphasizes the need to consider dataset-specific features when selecting an appropriate approach. These findings provide valuable insights into multimodal emotion recognition and pave the way for further research and applications in this domain.

Given the goal of deploying this model on autonomous machines, a crucial aspect for future research is exploring the inference process for minimal resource consumption. Identifying the lack of true “in the wild” datasets and developing a dataset that captures real-world scenarios with diverse expressions and environmental factors would be highly beneficial for evaluating the robustness of different approaches.

**Acknowledgements** This work is partially funded by the Spanish Ministry of Science and Innovation under project PID2021-122402OB-C22. By the ACIISI-Gobierno de Canarias and European FEDER funds under project ULPGC Facilities Net and Grant EIS 2021 04, it is also supported by “Programa Investigo” reference code 32/39/2022-0923131539 of Servicio Canario de Empleo. “Fondos del Plan de Recuperación, Transformación y Resiliencia - Next Generation EU”.

**Funding** Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

**Data Availability** The data supporting this research is available in various ways:

- **SAVEE**: Available upon request on: [SAVEE link](#).
- **CREMA-D**: Data deposited in an open repository at: [CREMA-D link](#).
- **RAVDESS**: Data deposited in an open repository at: [RAVDESS link](#).

## Declarations

**Conflicts of interest** The authors declare that they have no conflicts of interest.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

1. Aizi K, Ouslim M (2022) Score level fusion in multi-biometric identification based on zones of interest. *J King Saud Univ - Comput Inf Sci* 34(1):1498–1509. <https://doi.org/10.1016/j.jksuci.2019.09.003>
2. Bargshady G, Zhou X, Deo RC et al (2020) Enhanced deep learning algorithm development to detect pain intensity from facial expression images. *Expert Syst Appl* 149:113305. <https://doi.org/10.1016/j.eswa.2020.113305>
3. Bisogni C, Cimmino L, De Marsico M et al (2023) Emotion recognition at a distance: the robustness of machine learning based on hand-crafted facial features vs deep learning models. *Image Vis Comput* 136:104724. <https://doi.org/10.1016/j.imavis.2023.104724>
4. Busso C, Bulut M, Lee CC et al (2008) Iemocap: interactive emotional dyadic motion capture database. *Lang Resour Eval* 42(4):335–359. <https://doi.org/10.1007/s10579-008-9076-6>

5. Castro S, Hazarika D, Pérez-Rosas V et al (2019) Towards multimodal sarcasm detection (an \_obviously\_ perfect paper). In: Proceedings of the 57th annual meeting of the association for computational linguistics (vol 1, Long Papers). Association for Computational Linguistics, Florence, Italy
6. Cheng S, Zhou G (2020) Facial expression recognition method based on improved vgg convolutional neural network. *International J Pattern Recognit Artif Intell* 34(07):2056003. <https://doi.org/10.1142/S0218001420560030>
7. Choi JH, Lee JS (2019) Embracenet: a robust deep learning architecture for multimodal classification. *Inf Fusion* 51:259–270. <https://doi.org/10.1016/j.inffus.2019.02.010>
8. Chumachenko K, Iosifidis A, Gabbouj M (2022) Self-attention fusion for audiovisual emotion recognition with incomplete data. [arXiv:2201.11095](https://arxiv.org/abs/2201.11095)
9. Croitoru FA, Ristea NC, Ionescu RT et al (2022) Lerac: learning rate curriculum. [arXiv:2205.09180](https://arxiv.org/abs/2205.09180)
10. Deng J, Guo J, Zhou Y et al (2019) Retinaface: single-stage dense face localisation in the wild. [arXiv:1905.00641](https://arxiv.org/abs/1905.00641)
11. Do LN, Yang HJ, Nguyen HD et al (2021) Deep neural network-based fusion model for emotion recognition using visual data. *J Supercomput* 77(10):10773–10790. <https://doi.org/10.1007/s11227-021-03690-y>
12. Ekman P (1992) An argument for basic emotions. *Cognit & Emot* 6(3–4):169–200
13. Ekman P et al (1999) Basic emotions. *Handb Cognit Emot* 98(45–60):16
14. Fan H, Zhang X, Xu Y et al (2024) Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals. *Inf Fusion* 104:102161. <https://doi.org/10.1016/j.inffus.2023.102161>
15. Freire-Obregón D, De Marsico M, Barra P et al (2023) Zero-shot ear cross-dataset transfer for person recognition on mobile devices. *Pattern Recognit Lett* 166:143–150. <https://doi.org/10.1016/j.patrec.2023.01.012>
16. Freire-Obregón D, Hernández-Sosa D, Santana OJ et al (2023) Towards facial expression robustness in multi-scale wild environments. In: International conference on image analysis and processing
17. Haq S, Jackson P (2010) Machine audition: principles, algorithms and systems, IGI Global, Hershey PA, chap multimodal emotion recognition, pp 398–423
18. Haq S, Jackson P, Edge J (2008) Audio-visual feature selection and reduction for emotion classification. In: Proc Int Conf on auditory-visual speech processing (AVSP'08), Tangalooma, Australia
19. Kaya H, Gürpınar F, Salah AA (2017) Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image Vis Comput* 65:66–75. <https://doi.org/10.1016/j.imavis.2017.01.012>. <https://www.sciencedirect.com/science/article/pii/S0262885617300367>, multimodal Sentiment Analysis and Mining in the Wild Image and Vision Computing
20. Keutmann MK, Moore SL, Savitt A et al (2015) Generating an item pool for translational social cognition research: methodology and initial validation. *Behav Res Methods* 47(1):228–234
21. Khellat-Kihel S, Abrishambaf R, Monteiro J et al (2016) Multimodal fusion of the finger vein, fingerprint and the finger-knuckle-print using kernel fisher analysis. *Appl Soft Comput* 42:439–447. <https://doi.org/10.1016/j.asoc.2016.02.008>
22. Kim JY, Lee SH (2023) Coordvit: a novel method of improve vision transformer-based speech emotion recognition using coordinate information concatenate. In: 2023 International Conference on Electronics, Information, and Communication (ICEIC), pp 1–4. <https://doi.org/10.1109/ICEIC57457.2023.10049941>
23. Kumawat S, Verma M, Raman S (2019) Lbvcnn: local binary volume convolutional neural network for facial expression recognition from image sequences. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp 207–216. <https://doi.org/10.1109/CVPRW.2019.00030>
24. Li S, Deng W (2022) Deep facial expression recognition: a survey. *IEEE Trans Affect Comput* 13(3):1195–1215. <https://doi.org/10.1109/TAFFC.2020.2981446>
25. Lisetti C (1998) Affective computing -. *Pattern Anal Appl* 1:71–73. <https://doi.org/10.1007/BF01238028>
26. Liu S, He R (2024) Decision-level fusion detection method of hydrogen leakage in hydrogen supply system of fuel cell truck. *Fuel* 367:131455. <https://doi.org/10.1016/j.fuel.2024.131455>
27. Livingstone SR, Russo FA (2018) The ryerson audio-visual database of emotional speech and song (ravdess): a dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE* 13(5):1–35. <https://doi.org/10.1371/journal.pone.0196391>
28. Luna-Jiménez C, Kleinlein R, Griol D, et al (2022) A proposal for multimodal emotion recognition using aural transformers and action units on ravdess dataset. *Appl Sci* 12(1). <https://doi.org/10.3390/app12010327>
29. Middya AI, Nag B, Roy S (2022) Deep learning based multimodal emotion recognition using model-level fusion of audio-visual modalities. *Knowl-Based Syst* 244:108580. <https://doi.org/10.1016/j.knosys.2022.108580>



30. Pan B, Hirota K, Jia Z et al (2023) A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods. *Neurocomputing* 561:126866. <https://doi.org/10.1016/j.neucom.2023.126866>
31. Parkhi OM, Vedaldi A, Zisserman A (2015) Deep face recognition. In: *British machine vision conference*
32. Picard RW (1997) *Affective computing*. MIT Press, Cambridge, MA
33. Picard RW (2000) Toward computers that recognize and respond to user emotion. *IBM Syst J* 39(3.4):705–719. <https://doi.org/10.1147/sj.393.0705>
34. Poria S, Hazarika D, Majumder N et al (2019) MELD: a multimodal multi-party dataset for emotion recognition in conversations. In: Korhonen A, Traum D, Màrquez L (eds) *Proceedings of the 57th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, Florence, Italy, pp 527–536. <https://doi.org/10.18653/v1/P19-1050>
35. Ringeval F, Sonderegger A, Sauer JS et al (2013) Introducing the recola multimodal corpus of remote collaborative and affective interactions. In: *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)* pp 1–8
36. Russell JA (1983) Pancultural aspects of the human conceptual organization of emotions. *J Pers Soc Psychol* 45(6):1281
37. Ryumina E, Dresvyanskiy D, Karpov A (2022) In search of a robust facial expressions recognition model: a large-scale visual cross-corpus study. *Neurocomputing* 514:435–450. <https://doi.org/10.1016/j.neucom.2022.10.013>
38. Sadok S, Leglaive S, Séguier R (2023) A vector quantized masked autoencoder for speech emotion recognition. In: *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pp 1–5 <https://doi.org/10.1109/ICASSPW59220.2023.10193151>
39. Sajjad M, Ullah FUM, Ullah M et al (2023) A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines. *Alex Eng J* 68:817–840. <https://doi.org/10.1016/j.aej.2023.01.017>
40. Serengil SI, Ozpinar A (2020) Lightface: a hybrid deep face recognition framework. In: *2020 Innovations in intelligent systems and applications conference (ASYU)*, IEEE, pp 23–27 <https://doi.org/10.1109/ASYU50717.2020.9259802>
41. Shixin P, Kai C, Tian T et al (2022) An autoencoder-based feature level fusion for speech emotion recognition. *Digital Commun Netw*. <https://doi.org/10.1016/j.dcan.2022.10.018>
42. Singh M, Singh R, Ross A (2019) A comprehensive overview of biometric fusion. *Inf Fusion* 52:187–205. <https://doi.org/10.1016/j.inffus.2018.12.003>
43. Snyder D, Garcia-Romero D, Sell G et al (2018) X-vectors: robust dnn embeddings for speaker recognition. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp 5329–5333. <https://doi.org/10.1109/ICASSP.2018.8461375>
44. Vielzeuf V, Pateux S, Jurie F (2017) Temporal multimodal fusion for video emotion classification in the wild. In: *Proceedings of the 19th ACM international conference on multimodal interaction*. Association for Computing Machinery, New York, NY, USA, ICMI '17, pp 569–576. <https://doi.org/10.1145/3136755.3143011>
45. Viola P, Jones MJ (2004) Robust real-time face detection. *Inte J Comput Vis* 57(2):137–154. <https://doi.org/10.1023/B:VISI.0000013087.49260.fb>
46. Vu MT, Beurton-Aimar M, Marchand S (2021) Multitask multi-database emotion recognition. In: *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp 3630–3637. <https://doi.org/10.1109/ICCVW54120.2021.00406>
47. Wang S, Zheng Z, Yin S et al (2020) A novel dynamic model capturing spatial and temporal patterns for facial expression analysis. *IEEE Trans Pattern Anal Mach Intell* 42(9):2082–2095. <https://doi.org/10.1109/TPAMI.2019.2911937>
48. Wang W, Li Q, Xie J et al (2023) Research on emotional semantic retrieval of attention mechanism oriented to audio-visual synesthesia. *Neurocomputing* 519:194–204. <https://doi.org/10.1016/j.neucom.2022.11.036>
49. Xiang J, Zhu G (2017) Joint face detection and facial expression recognition with mtcn. In: *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*, pp 424–427. <https://doi.org/10.1109/ICISCE.2017.95>
50. Zhang Z, Luo P, Loy CC et al (2018) From facial expression recognition to interpersonal relation prediction. *Int J Compuy Vis* 126(5):550–569. <https://doi.org/10.1007/s11263-017-1055-1>