Knowledge-Based Systems 244 (2022) 108580

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities

Asif Iqbal Middya [a], Baibhav Nag [b], Sarbani Roy [a,*]

[a] Department of Computer Science and Engineering, Jadavpur University, Kolkata, India
[b] Department of Mathematics, Jadavpur University, Kolkata, India

ARTICLE INFO

Article history:
Received 8 October 2021
Received in revised form 31 January 2022
Accepted 9 March 2022
Available online 16 March 2022

Keywords:
Multimodal emotion recognition
Audio features
Video features
Classification
Deep learning

ABSTRACT

Emotion identification based on multimodal data (e.g., audio, video, text, etc.) is one of the most demanding and important research fields, with various uses. In this context, this research work has conducted a rigorous exploration of model-level fusion to find out the optimal multimodal model for emotion recognition using audio and video modalities. More specifically, separate novel feature extractor networks for audio and video data are proposed. After that, an optimal multimodal emotion recognition model is created by fusing audio and video features at the model level. The performances of the proposed models are assessed based on two benchmark multimodal datasets namely Ryerson Audio–Visual Database of Emotional Speech and Song (RAVDESS) and Surrey Audio–Visual Expressed Emotion (SAVEE) using various performance metrics. The proposed models achieve high predictive accuracies of 99% and 86% on the SAVEE and RAVDESS datasets, respectively. The effectiveness of the models are also verified by comparing their performances with the existing emotion recognition models. Some case studies are also conducted to explore the model's ability to capture the variability of emotional states of the speakers in publicly available real-world audio–visual media.

1. Introduction

Human communication is heavily influenced by their emotional states. Identification of the emotional states is a challenging task in many fields with multiple applications [1–3] including lie detection, audio–visual surveillance, affective computing, online teaching–learning, online meeting, human–computer interaction (HCI), and many more. The study of emotional variability is an important factor for investigating psychological adaptation and well-being. Moreover, it is also important for machines to be able to recognize human emotions in order to make better decisions. It is because intelligent machines have become an indispensable part of our daily lives, hence developing methods to help them correctly classify users' emotions has become essential for the advancement of society.

There are a variety of emotional models [4,5] in the literature, but two of them are commonly used in research: dimensional models and discrete emotion models. The dimensional models represent emotions as a continuous spectrum while the latter one describes emotions as discrete values. Both models are extensively adopted by psychologists for analyzing emotions. Human emotion recognition can be accomplished through a variety of means, including speech data, facial expressions, body gestures, physiological parameters, and many others. Because each of these modalities is distinct, fusing their results in a rich representation of features capable of performing emotion recognition efficiently. Previous research works have shown that relying solely on a single modality for emotion recognition is inefficient [6,7]. More specifically, some existing literature also demonstrates that using multiple modalities (audio, video, text, etc.) for emotion recognition yields significantly better results than using only one [8].

In the above-mentioned context, multimodal real-time analysis of emotional states has recently received huge attention worldwide. Dynamic multimodal analysis outperforms static analysis of human emotions as it uses features like changes in eye movements and facial expressions over time and also considers the characteristics of speech [9]. However, in this area of research, finding an efficient multimodal model that is not computationally heavy and overly complex remains a challenging task. Hence, in this work, an efficient and relatively lightweight multimodal fusion model that considers both audio and video data is proposed to identify emotional states. This work's overall contributions can be described as follows:

- Deep learning-based feature extractor networks for video and audio data are proposed. A model-level fusion of the video and audio features is performed to create an optimal multimodal emotion recognition model.

* Corresponding author.
E-mail addresses: asifim.rs@jadavpuruniversity.in (A.I. Middya), nagbaibhav@gmail.com (B. Nag), sarbani.roy@jadavpuruniversity.in (S. Roy).

https://doi.org/10.1016/j.knosys.2022.108580
0950-7051/

Table 1
A summary of emotion categorization models and algorithm.

Author	Models	#Emotions	Emotion type	Emotions
Ortony et al. [12]	OCC	22	Negative	Pity, resentment, disappointment, fears-firmed, hate, reproach, shame, distress, fear.
			Positive	Gloating, happy, relief, satisfaction, gratitude, gratification, love, admiration, pride, joy, hope
Shaver et al. [13]	–	6	Negative	Fear, anger, sadness.
			Positive	Surprise, love, joy.
Cambria et al. [14]	Hourglass model	24	Aptitude	Loathing, disgust, boredom, acceptance, trust, admiration.
			Sensitivity	Terror, fear, apprehension, annoyance, anger, rage.
			Attention	Amazement, surprise, distraction, interest, anticipation, vigilance
			Pleasantness	Grief, sadness, pensiveness, serenity, joy, ecstasy
P. Ekman [15]	–	6	Negative	Disgust, fear, anger, sadness.
			Positive	Surprise, joy.
A. T. Latinjak [16]	Latinjak cube	20	Neutral	Fatigued, surprise, calm, alerted.
			Positive	Over-confident, relaxed, relieved, optimistic, satisfied, excited, enjoying, elated.
			Negative	Dejected, bored, sad, pessimistic, deceived, anxious, distressed, angry.
R. Plutchik [17]	Wheel of emotions	8	Negative	Disgust, fear, anger, sadness.
			Positive	Anticipation, surprise, trust, joy.

- Rigorous experimental analyses are carried out to assess the performance of the presented multimodal emotion recognition model on two benchmark databases namely SAVEE (Surrey Audio–Visual Expressed Emotion) [10] and RAVDESS (Ryerson Audio–Visual Database of Emotional Speech and Song) [11] containing audio–visual data. The evaluation results also indicate that the proposed model achieves better predictive performances as compared to the existing multimodal emotion recognition models. Additionally, case studies are conducted for exploring the variability of emotional states of the subjects in various publicly available audio–visual media.

The remainder of the paper is structured as given below. A comprehensive review of the existing works is presented in Section 2. The methodology of the work includes datasets, data preparation, feature extraction, and model building. The experimentation and the performance analysis are provided in Section 4. The conclusion and future scope are provided in Section 5.

2. Related works

A thorough investigation of traditional and deep learning-based methods for multimodal emotion recognition is provided in this section. Because of its wide range of applications, multimodal emotion classification has gained the attention of researchers all over the world, and a significant amount of research is being done in this area each year.

In the literature, researchers and psychologists have proposed a variety of emotion categorization models [12–17,35]. Shaver et al. [13] devised one of the earliest emotion modeling techniques. The authors categorized emotions into prototypes assuming that various portions of emotional information build up an organized whole. In [36], the authors proposed that all emotions are distinct, self-contained, and linked to one another via a hierarchical form. Ortony et at. [12] devised an emotion model called OCC (named after the initials of the 3 authors — Ortony, Clore, and Collins) where related emotions are categorized depending on their arousal strengths. Steunebrink et al. [37] went over the flaws of the OCC model in greater detail. They updated the OCC model to include two different emotions: "disgust" and "interest". Another interesting model is presented by Ekman [15], in which emotions are viewed as quantifiable and physiologically distinguishable. Ekman introduced a total of six emotions namely surprise, sadness, joy, disgust, fear, and anger. The Ekmann model was modified by Plutchik into a "Wheel of Emotions" model [17]. Multi-dimensional models are also introduced as emotion models. The Latinjak cube 3-dimensional model [16] is a notable example of a 3-dimensional emotion model. It can accommodate twenty different emotions into the 3-dimensional mode by using the 3 dimensions. Finally, Cambria et al. [14] introduced the "Hourglass model" a psychologically and biologically inspired novel emotion classification model which is based on the aforementioned models. Attention, pleasantness, aptitude, and sensitivity are the 4 affective dimensions in the Hourglass model. A summary of emotion categorization models and algorithms is provided in Table 1.

Several studies apply traditional methods for multimodal emotion recognition. At the decision and feature levels, Busso et al. [33] demonstrated the benefits and drawbacks of an emotion recognition system based on visual expression or acoustic information, as well as the use of two modalities in combination. The audio–visual data was used to classify the emotional states based on the support vector machine, demonstrating that the system performed better than either of the unimodal systems. The overall classification accuracy of the feature-level fusion approach was 89.1% while for the decision-level fusion it was 89.0%. Wang et al. [32] introduced an emotion detection approach using audio–visual modalities. They extracted the facial features using a Gabor filter bank and used various speech features. Using Fisher's Linear Discriminant Analysis (FLDA) approach, the step-wise strategy was utilized to choose the key features for the classification of various emotions. The proposed system achieved an overall accuracy of 82%. Yan et al. [28] demonstrated an innovative bimodal emotion identification method that incorporates both face and speech features. Xu et al. [27] demonstrated emotion recognition using a combination of human voices and facial expressions. They used a dataset called CHEAVD [38], which contains eight emotions. SVM was employed to identify the facial and speech features. The classification outcomes were then combined using Bayesian rules at the decision level. The overall performance of their model yielded an accuracy of 38% for the CHEAVD dataset. Rao et al. [23] use the SVM model to perform decision level fusion for the detection of several emotional states. Yoshitomi et al. [34] presented a multimodal approach that takes into account not only audio and visual data but also the thermal distributions captured by an infrared camera. They used experimentally determined weights to integrate these modalities at the decision-level.

Deep learning approaches have recently outperformed traditional methods in many domains of data science. As a result, it is no surprise that they have been adopted here as well. Nguyen et al. [25] introduced a novel method that integrated 3D CNN for

Table 2
Summary of proposed multimodal emotion recognition approaches in recent decades.

Author	Models	Description of methods
Bagadi et al. [18]	Long short-term memory (LSTM)	The RAVDESS and SAVEE datasets are used for multimodal emotion identification utilizing text and audio.
Abdullah et al. [19]	Convolutional neural network (CNN) + recurrent neural network (RNN)	Performed facial expression identification on RAVDESS video files by using CNN + RNN to capture temporal as well spatial features.
Krishna et al. [20]	Used 1D CNN and attention mechanism to leverage data from both audio and textual cues	Proposed a novel approach based on CNN with cross-modal focus for emotion classification.
Jaratrotkamjorn et al. [21]	DBN (Deep Belief Network)	Extracted audio–visual features and fused them at feature-level and used them for emotion classification.
Sahu [22]	Used both classical classifiers as well as neural network and LSTM models	It was demonstrated that shallow machine learning techniques trained on the hand-crafted features can reach comparable performance to the deep learning approaches.
Prasada Rao et al. [23]	Support vector machines (SVM)	Performed decision level fusion using SVM for emotion classification.
Yoon et al. [24]	Dual Recurrent Neural Networks (RNNs)	Encoded information from both audio and textual cues and then combines them for emotion classification.
Nguyen et al. [25]	3D CNN + Deep Belief Networks (DBNs)	Extracted audio–visual features using the mentioned approaches and performed emotion recognition.
Miao et al. [26]	Classical (SVM, REPTree, Random Forest) as well as deep learning models (DBN, RNN)	Did multimodal emotion recognition from facial and acoustic signals.
Xu et al [27]	SVM	Classified audio–visual features using SVM and then integrated the classification results at the decision level using Bayesian rules.
Yan et al. [28]	SVM and SR (Sparse Representation) approaches	Used SKRRR (Sparse Kernel Reduced Rank Regression) method to fuse audio and video data and employed SVM and SR to perform emotion recognition.
Kahou et al. [29]	Ensemble of deep learning models	Performed effective emotion recognition using ensemble learning.
Srivastava et al. [30]	Multimodal Deep Boltzmann Machines	Employed multimodal deep belief networks that outperformed SVM.
Ngiam et al. [31]	Sparse Restricted Boltzmann Machines (RBMs)	Performed cross-modal and shared representation learning and multimodal fusion for effective emotion classification.
Wang et al. [32]	Fisher's Linear Discriminant Analysis (FLDA) classifier	Facial features and the MFCC features of the audio are merged into a single long feature vector and considered as input to the classifier.
Busso et al. [33]	SVM	They performed decision as well as feature level fusion of audio–visual data and showed that bimodal emotion recognition is better than its unimodal counterpart.
Yoshitomi et al. [34]	HMM (Hidden Markov Model) and neural networks	Proposed a multimodal system that combines audio–visual and thermal information obtained from infrared cameras at the decision level.

multimodal emotion detection. On the eNTERFACE database [39], it was employed to classify human emotions at the score level fusion, with an accuracy of 89 percent. Miao et al. [26] presented a CNN-based multimodal emotional state detection system from facial expressions and acoustic information. Ngiam et al. [31] showed impressive results on audio–visual speech classification. On the CUAVE [40] and AVLetters [41] datasets, they used Restricted Boltzmann Machines (RBM), representation learning, and multimodal fusion. To detect emotional states from video clips, Kahou et al. [29] implemented an ensemble of deep learning techniques. The work has been chosen as the winner of the Emotion Recognition in the Wild Challenge [42]. Krishna et al. [20] performed emotion recognition from audio and textual information. To achieve an improved emotion detection model, they employed raw audio analysis with 1D convolutional models and attention mechanisms across audio and text information. Multimodal emotion recognition has been recently enhanced using transformer architecture. They learn emotional temporal dynamic information through a self-attention mechanism. Several existing works highlight the significance of multimodal transformers for multimodal emotion recognition [43,44]. For instance, Huang et al. [43] utilized a transformer for a model-level fusion of the audio–visual modalities for emotion recognition. They used the transformer in conjunction with an LSTM to enhance their arousal and valence CCCs (Concordance Correlation Coefficients) for both unimodal and multimodal data. A summary of the related works in this domain has been provided in Table 2.

Moreover, regarding multimodal emotion detection, there are also several recent literature in the field of affective computing [45–47], sentiment analysis [48–50] and reinforcement learning based emotion recognition [51]. In [48], Cambria et al. proposed a sentic blending approach intertwined with a facial emotion classification model and a sentiment mining engine for continuous evaluation of semantics and sentics on facial expression datasets. Chaturvedi et al. [49] implemented a multimodal sentiment detection model for sentiment analysis on 4 benchmark datasets-Trip advisor text dataset, multimodal opinion utterances dataset (MOUD), amigos dataset and the multi-domain sentiment analysis dataset. However, in contrast to our proposed work, they utilized a convolutional fuzzy sentiment classifier. Stappen et al. [50] utilized a lexical knowledge-based extraction approach for topic understanding and sentiment analysis of videos of the MuSe-CAR dataset. They attempted to investigate the video content and estimate the emotional valence, arousal and speaker topic categories. Zhang et al. [51] proposed a new multimodal emotion recognition model named ERLDK based on reinforcement learning and domain knowledge for emotion detection of conversational videos on the IEMOCAP and MELD datasets. In contrary to Stappen et al. [50] and Zhang et al. [51] we have attempted to identify emotion based on both video as well as audio features. Susanto et al. [45] in his work re-examined an emotion classification technique called "The Hourglass of Emotions" optimized for polarity identification. Finally [46,47] respectively propose OntoSenticNet and a modified version of this model for effective sentiment analysis from multimodal resources.

The majority of the existing research works have employed classical or deep learning-based approaches to demonstrate that multimodal emotion recognition is superior to unimodal one. They have employed either decision-level, feature-level or model-level fusion to achieve their targets. In contrast to previous studies, this study conducted extensive research on the fusion part to see whether the model's performance could be improved even further. Specifically, in this study, the model-level fusion of audio and video modalities is investigated in-depth in order to determine the optimal multimodal model that will yield the best classification result. The sensitivity of the model parameters is also analyzed in order to confirm the model's robustness. Compared to the existing deep learning-based approaches, the proposed models are lightweight, simple and efficient and consist of only 6,028,104 and 3,037,063 parameters respectively. Note that these numbers of parameters are much smaller than the typical large deep learning-based models proposed in recent times such as VGG-16 [52] and Inception-ResNet-v2 [53]. For instance, Inception-ResNet-v2 and VGG-16 have 55.8 million and 138 million parameters respectively and thus they are extremely memory-intensive.

3. Methodology

A brief overview of the datasets used in this study, the data preprocessing, and feature extraction are presented in this section. An elaborated description is also provided for the proposed models and the methodology of work.

3.1. Datasets

Two benchmark multimodal databases, the SAVEE [10] and the RAVDESS [11] are used in this study.

RAVDESS [11] is a multimodal dataset having 7356 files (24.8 GB). Data from 24 professional actors (12 male, 12 female), speaking two lexically-matched sentences in North American accent are included here. Speech exhibits disgust, surprise, fearful, sad, happy, angry, and calm emotions. There are different levels of emotional intensities for every expression of emotion (strong and normal), as well as a neutral expression.

Video-only (no sound), Audio–Video, and Audio-only are the modalities of data contained in this database. A portion of the video speech files (i.e., files with both audio and visual modalities) is used because the goal of this study is to perform multimodal emotion recognition, which necessitates the use of both audio and video data for each actor. There are 1440 files in total, which are divided into eight emotion classes: fearful, disgust, angry, sad, happy, calm, neutral, and surprised. Fig. 1 shows a sample waveform for different emotional states in this dataset. On the other hand, Fig. 2 provides the images of facial expressions of each emotion for the various actors.

Four researchers and students from the University of Surrey, ranging in age from 27 to 31, recorded the data for the SAVEE [10] dataset. This dataset contains 480 audio–visual files, with 120 utterances for each speaker. All the audio–visual files are in .avi format and there are seven emotion classes namely surprise, sadness, neutral, happiness, fear, disgust, and anger.

3.2. Data preprocessing and feature engineering

In this section we describe how we extracted our video and audio features and prepared our multimodal labeled dataset.

3.2.1. Video preprocessing and feature engineering

A total of 6 frames per video (duration = 3 s) is extracted for every actor of each dataset as it is found that 6 frames would be enough to capture the spatio-temporal information associated with a particular emotion. Note that the frames from the videos are extracted using the computer vision library OpenCV (version 3.3) [54]. The inbuilt pre-trained deep learning-based face detector of OpenCV is then employed for accurate facial feature extraction from the frames. Specifically, Caffe-based face detector is used in this work. The model's architecture and the weights of each layer are stored in separate `.prototxt` and `.caffemodel` files that are downloaded and used in the runtime. The deep learning-based face detector uses the Single Shot Detector (SSD) framework. After extracting the frames from each video, they are scaled down to 40% of their original size. Each image is then passed through the inbuilt `blobFromImage()` function of the DNN module to create a blob. The `blobFromImage()` takes care of the necessary image preprocessing steps like setting the blob dimensions and RGB normalization. It returns a blob which is a 4-dimensional representation of the input image after mean subtraction of pixel intensities and normalization. Now, to detect faces, the blob is passed through the network and the detection scores (probabilities) of the facial predictions are found. The detection scores are compared with the confidence threshold in order to filter out the weak detections. After that a bounding box along the facial area is identified by computing x-y coordinates of the box from the detection scores and the image is cropped accordingly. Finally, the cropped image will be resized into 64×64 pixels and append to a list of features. As a result, facial features from each video are extracted and stored in the video features list in this manner.

3.2.2. Audio preprocessing and feature engineering

A Python library called `moviepy` is employed for extracting the audio content from each video. Since the audio extracted was found to be a stereo file, it is split into a mono one using the `pydub` library [55]. then feature extraction is performed from the mono audio file. Various audio features namely MFCC, melspectrogram, spectral contrast and tonnetz are extracted and stored in the audio feature list. The significance of each of these features is discussed below.

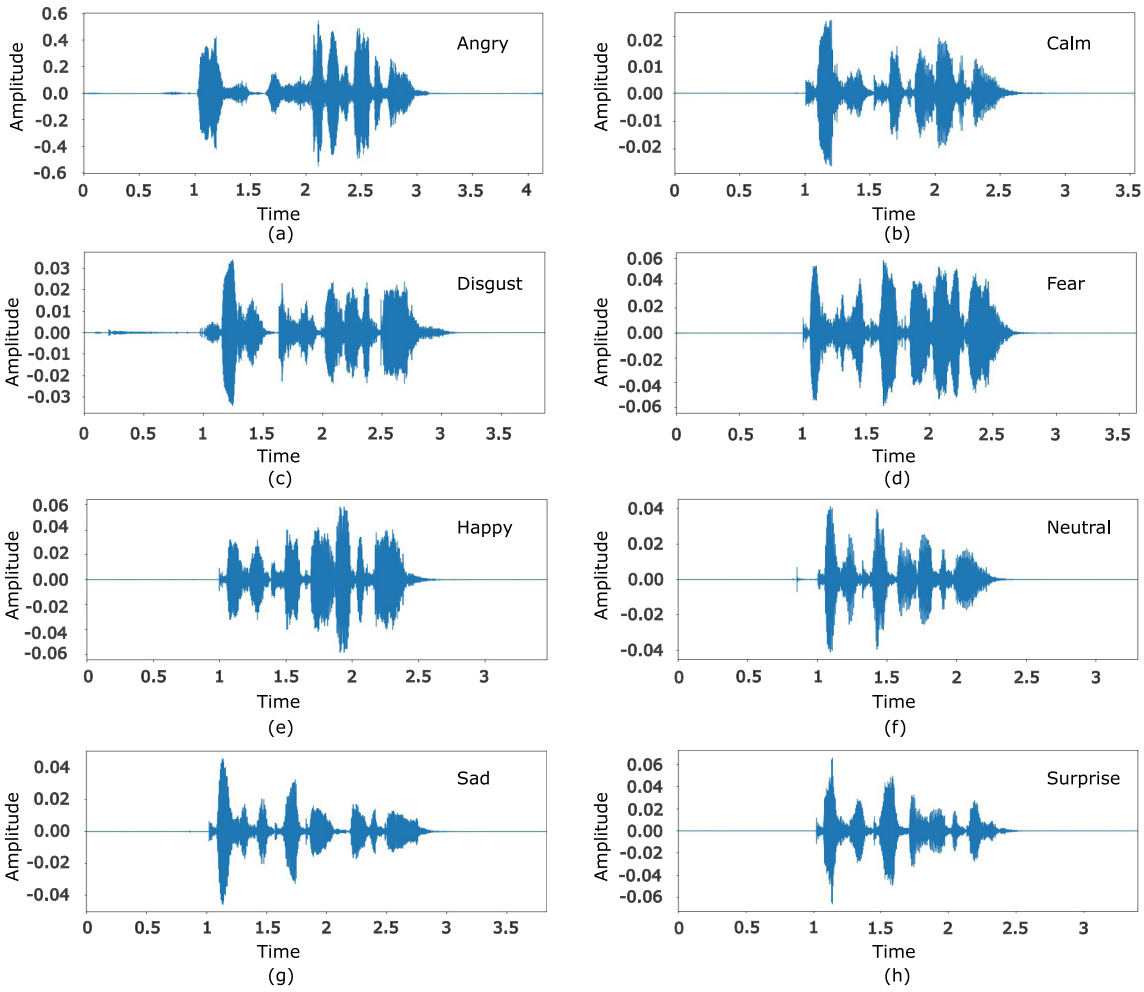

Fig. 1. Images of waveform for various types of emotion file of RAVDEES dataset (a) angry (b) calm (c) disgust (d) fear (e) happy (f) neutral (g) sad (h) surprise.

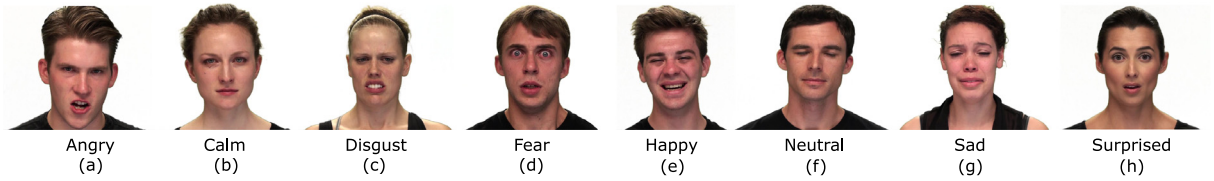

Fig. 2. Sample video frames of different facial expressions for each emotion by various actors of RAVDESS dataset (a) angry (b) calm (c) disgust (d) fear (e) happy (f) neutral (g) sad (h) surprise.

MFCCs and melspectrograms have extensive applications in the field of audio analysis [56]. They have been considered as important features for emotion identification from audio. Note that DCT type-2 [57] is employed for extracting 40 MFCCs. The mathematical form of DCT Type-2 is as follows.

$$X_j = \sum_{n=0}^{M-1} x(n) \cos\left(\frac{\pi .(n+\frac{1}{2}).j}{M}\right) \quad (1)$$

here $x(n)$ represents the signal stream, M denotes the total number of elements and $0 \leq j \leq M\text{-}1$, *cos* is the trigonometric cosine function, π is a constant. Now, spectral contrast gives a comprehensive spectral depiction of audio. According to the literature [58], methods that rely on detailed spectral contrast features show better performance than those that rely solely on mel-features, such as MFCCs, and hence they are used here as well.

Tonnetz is a conceptual lattice diagram that depicts tonal space. In terms of harmony and pitch classes, it is analogous to chromagram.

3.3. Proposed models

Two benchmark datasets (e.g., SAVEE and RAVDESS) are used in this work for building a multimodal fusion model for emotion recognition. However, there is no generic model found which outperforms both datasets simultaneously. Separate deep learning-based feature extractor networks are proposed for different modalities.

3.3.1. Model building based on RAVDESS dataset

Several CNN-based audio and video feature extractor networks are devised to find out the best combination of the feature extractor networks that could be useful to build the optimal multimodal

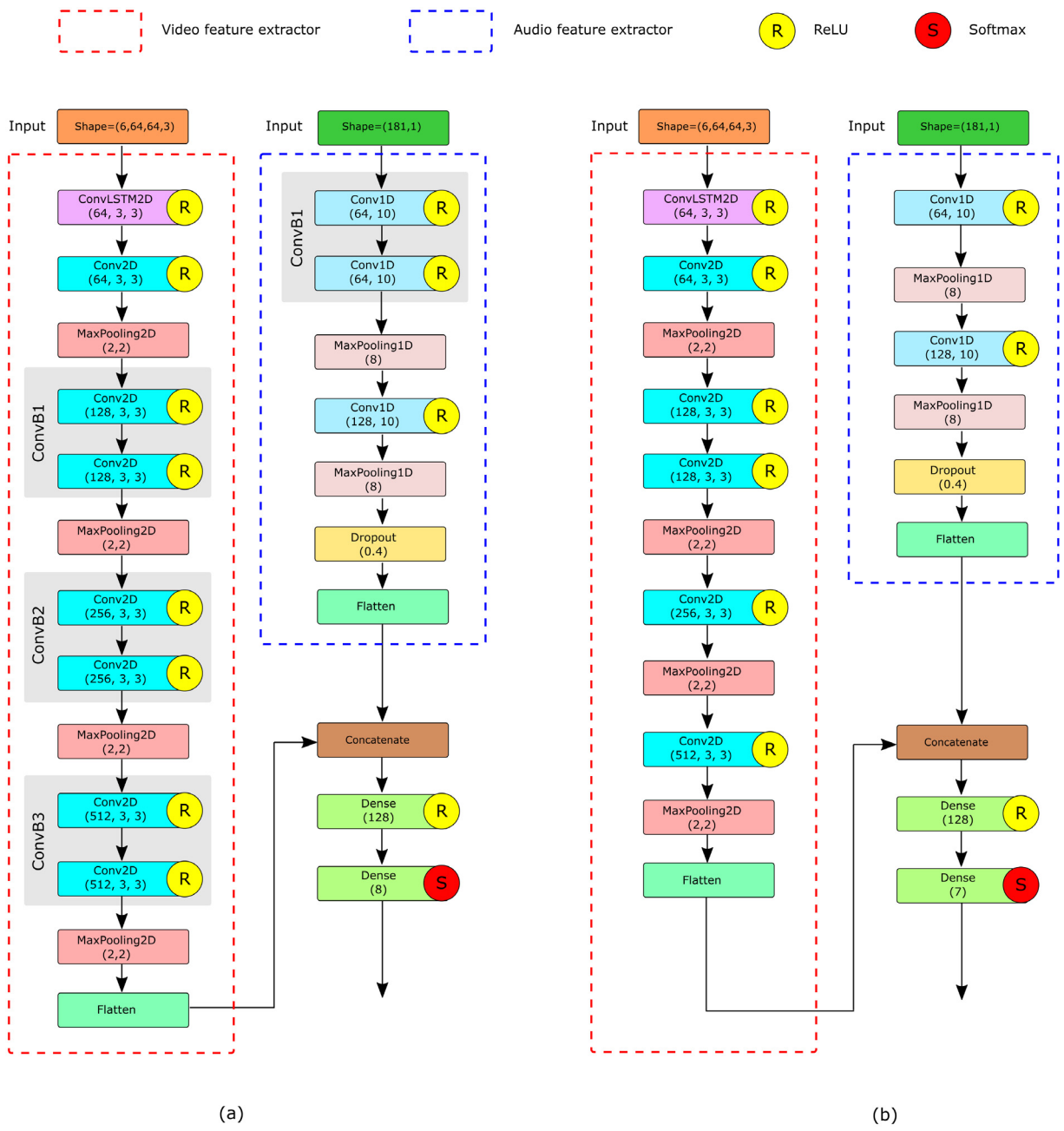

Fig. 3. Architecture of the proposed models for (a) RAVDESS dataset with the combination V8+A4 (b) SAVEE dataset with the combination V2+A2.

model. In other words, a model-level fusion of the audio–visual modalities is performed in this work. Note that these audio and video feature extractor networks are essentially variants of each other with respect to their number of convolution (C layer) and pooling (P layer). As shown in Table 3, a total of 9 video feature extractor networks (V1 to V9) and 6 audio feature extractor networks (A1 to A6) are investigated based on different combinations of convolution and pooling layers. Hence, there can be a total of 54 possible combinations of these 9 video and 6 audio feature extractor networks. Here, the aim is to conduct a rigorous and thorough investigation of model-level fusion of audio–visual modalities to find out the optimal multimodal model for effective human emotion recognition. Note that the number of model-level combinations in the exploration space is not reasonably high. Thus it is computationally feasible to adapt a brute force approach here to check all the possibilities for obtaining the optimal multimodal model. This method always guarantees to find out the best solution (i.e., the optimal model) among all the entries in the chosen exploration space and is generic and not limited to any specific domain of problems. Upon experimenting (see Section 4 for more details), it is found that the combination V8+A4 yield better result compared to the others. The architecture of the proposed model with the combination V8+A4 as the feature extractors is provided in Fig. 3(a). The details of the overall model architecture are discussed below.

- *Video feature extractor (V8)*: As shown in Fig. 3(a), a 2-dimensional CNN has been introduced as the video feature extractor network consisting of convolution, pooling and flatten layers. This part is shown in a red dashed line rectangular box. The input shape is of size (6, 64, 64, 3) where 6 is the number of frames extracted per video, 64 × 64 is the frame size and 3 is the number of channels (Red, Green, Blue). The input is

Table 3
Video and audio feature extractor models with different number convolution and pooling layers.

Modality	Feature extractor	#Convolution layer	#Pooling layer
Video	V1	6	5
	V2	6	4
	V3	6	3
	V4	7	5
	V5	7	4
	V6	7	3
	V7	8	5
	V8	8	4
	V9	8	3
Audio	A1	2	1
	A2	2	2
	A3	3	1
	A4	3	2
	A5	4	1
	A6	4	2

fed to the first layer which is a ConvLSTM2D layer of 64 filters with a kernel size of 3 × 3. The ConvLSTM2D has proven to be very powerful in extracting spatio-temporal features from images that change with time. Since the frames have been fed as input, it is imperative that the model must understand how the facial expressions change over time sequentially and thus extract the necessary features for effective emotion classification. The output of the ConvLSTM2D layer is fed to a Conv2D layer having 64 filters. The main differences between Conv2D and ConvLSTM2D are as follows. Conv2D or 2-dimensional convolutional neural network layer is used to deal with image features that are not sequential or temporal in nature. On the other hand, the ConvLSTM2D is basically a merger of 2-dimensional convolution layer and LSTM layer used for dealing with image features that have a temporal dimension. Its output is then fed to a max-pool layer of window size 2 × 2 with a stride of 2. In the remaining part, there are 3 convolutional blocks (ConvB1, ConvB2, and ConvB3) each having 2 conv2D layers. The convolutional blocks ConvB1, ConvB2, and ConvB3 have 128, 256, and 512 filters with a kernel size of 3 × 3. All the layers have the ReLU function as their activation function. The ReLU function is a widely used activation function in deep learning models for its effective results. It is a linear function whose range lies in [0, ∞] and is given by the below equation:

$$F(a) = max(0, a) \tag{2}$$

where a represents the input to a neuron and $max(\cdot)$ will return the maximum of 0 and a. A max-pool layer with window size 2 × 2 follows each of the convolutional blocks. Finally, the output is flattened after the last pooling layer.

- *Audio feature extractor (A4)*: The feature extractor for the audio model is a 1-dimensional CNN having convolution, pooling, dropout and flatten layers. This audio feature extractor is shown in a blue dashed line rectangular box. It consists of a total of 3 convolution and 2 pooling layers. As shown in Fig. 3(a), the input shape for audio feature extractor is (181 × 1) where 181 is the total feature set size (mfccs + melspectrograms + spectral contrast + tonnetz). There are 64 filters in the first two convolutional layers, with a kernel size of 10 × 1 and with same kernel size stride of 1. After that, there is a max-pool layer of window size 8. The next layer is a convolution layer with 128 filters, a kernel size of 10 × 1, and a stride of 1. Here, a l2 regularization of 0.01 is employed as both our kernel and bias regularization. The output is then fed to a max-pooling layer with size 8 followed by a dropout of 0.4. All the layers have the ReLU activation function. Finally, the output received from the pooling layer is flattened.
- *Fusion*: Two flattened outputs received from video and audio feature extractors are concatenated for classifying the emotional states. In Fig. 3, *concatenate* refers to the layer that is used to merge flattened outputs from video and audio feature extractors into a single long feature vector. The concatenate layer's output is sent into a dense layer having 128 filters and an activation ReLU. The last layer is a fully-connected layer having 8 units for the 8 emotion classes. It uses the softmax activation function for the final emotion classification. The softmax function is expressed as follows.

$$\sigma(q)_i = \frac{e^{q_i}}{\sum_{k=1}^{\delta} e^{q_k}} \tag{3}$$

The function uses exponential function to every element q_i of vector $q = (q_1, q_2, q_3, \ldots, q_\delta)$ and normalizes the values. Here, δ indicates the number of emotion classes. The Adam optimizer [59] having a learning rate of 0.0005 and a decay rate of 1e-07 is used. The categorical cross-entropy loss is considered as the loss function and the accuracy metric is used as our evaluation metric. The categorical cross-entropy loss is widely utilized in multi-class classification.

3.3.2. Model building based on SAVEE dataset

The same 54 combinations of audio and video feature extractors are also implemented here that are previously applied for the RAVDESS dataset. The aim here is to see if the combination V8+A4 gives the best result here as well, that is, we wanted to see whether the multimodal model is a generic model or dataset-specific. On experimenting (see Section 4 for more details), it is found that the model is dataset-specific. The V2+A2 combination achieves the best outcome here for the SAVEE dataset. The detailed architecture of the model with the combination V2+A2 is presented in Fig. 3(b).

- *Video feature extractor (V2)*: As shown in Fig. 3(b), a 2-dimensional CNN has also been used here as the video feature extractor network consisting of convolution, pooling and flatten layers. The input shape is the same as earlier, i.e., (6, 64, 64, 3). The model has a total of 6 convolution and 4 pooling layers. The input is fed to a layer which is again a ConvLSTM2D layer of 64 kernels each of size 3 × 3. This layer's output is sent to another convolution layer of 64 filters, kernel size of 3 × 3. Its output is then fed to a max-pool layer of size 2 × 2. Next, there are two consecutive convolution layers of 128 filters each, kernel size of 3 × 3 followed by a pooling layer. Next, there are two convolution layers having 256 and 512 filters, and each of them is followed by a pooling layer. The convolution layers have ReLU as their activation function given by Eq. (2). Finally, the output of our last pooling layer is flattened.
- *Audio feature extractor (A2)*: The same input shape (i.e., 181 × 1) is used in the audio feature extractor (A2) which is a 1-dimensional CNN like before having 2 convolution and 2 pooling layers. The first layer has

Table 4
Performance (in terms of classification accuracy) for different combinations of audio and video feature extractor networks for RAVDESS and SAVEE datasets.

Dataset	Audio feature extractors	Video feature extractors								
		V1	V2	V3	V4	V5	V6	V7	V8	V9
RAVDESS	A1	0.76	0.75	0.73	0.75	0.74	0.74	0.74	0.72	0.76
	A2	0.54	0.55	0.60	0.56	0.60	0.55	0.54	0.58	0.51
	A3	0.75	0.77	0.74	0.75	0.76	0.72	0.74	0.74	0.76
	A4	0.69	0.71	0.68	0.69	0.71	0.71	0.66	0.86	0.66
	A5	0.71	0.76	0.72	0.74	0.70	0.76	0.76	0.74	0.74
	A6	0.73	0.63	0.72	0.67	0.63	0.63	0.75	0.69	0.67
SAVEE	A1	0.94	0.73	0.67	0.63	0.67	0.63	0.73	0.69	0.65
	A2	0.88	0.99	0.94	0.98	0.60	0.65	0.63	0.92	0.52
	A3	0.69	0.73	0.75	0.69	0.73	0.73	0.69	0.79	0.69
	A4	0.98	0.98	0.94	0.60	0.75	0.58	0.73	0.69	0.60
	A5	0.88	0.90	0.67	0.96	0.71	0.69	0.71	0.75	0.75
	A6	0.85	0.96	0.90	0.71	0.92	0.90	0.65	0.58	0.69

Table 5
A detailed performance of the proposed models (i.e., model with V8 + A4 for RAVDESS and V2 + A2 for SAVEE) along with some other randomly chosen well-performing combinations of audio and video feature extractors.

Dataset	Best performing combinations	Accuracy	Precision	Recall	F1-score	Specificity	AUC
RAVDESS	V8 + A4	0.86	0.86	0.86	0.86	0.98	0.98
	V5 + A3	0.76	0.76	0.75	0.75	0.97	0.96
	V6 + A5	0.76	0.76	0.75	0.75	0.97	0.96
	V9 + A3	0.76	0.77	0.75	0.75	0.97	0.96
SAVEE	A2 + V2	0.99	0.99	0.99	0.99	1.00	1.00
	V1 + A4	0.98	0.98	0.98	0.98	1.00	1.00
	V2 + A4	0.98	0.99	0.98	0.98	1.00	1.00
	V4 + A2	0.98	0.97	0.95	0.96	1.00	0.99

64 filters having a kernel size of 10×1 and stride of 1. Its output is fed to a max-pool layer of window size 8. Next, there is a convolution layer having 128 filters with the same kernel size and stride. Here we have employed a l2 regularization of 0.01 as both our kernel and bias regularization. The output is fed to a max-pool layer of size 8 followed by a dropout of 0.4. All our layers have the ReLU activation function. Finally, we flatten the output received from the pooling layer.

- *Fusion*: The fusion part is identical to what is used in the model architecture for the RAVDESS dataset. The only difference here is that the final fully connected layer here has 7 units as there are 7 emotion classes in the SAVEE dataset. The rest of the parameters like the optimizer, decay rate, loss function and evaluation metric are the same.

4. Experimentation and results

This section uses rigorous analysis to evaluate the efficacy of the proposed multimodal models. All the models have been implemented with the Adam optimizer [59] having a learning rate of 0.0005 and a batch size of 64. For the model training, a total of 100 epochs were considered. The train, validation, and test set contains 80%, 10%, and 10% data respectively. All the experiments have been conducted on Google Colaboratory notebooks with Intel (R) Xeon (R) CPU @ 2.00 GHz, 39 MB cache, and NVIDIA Tesla T4 GPU (CUDA version: 11.2), 13 GB usable memory. The models are implemented in Python 3.7.10 with the following python libraries: librosa (version: 0.8.1), scikit-learn (version: 0.22.2), tensorflow (version: 2.6.0), keras (version: 2.6.0), pandas (version: 1.1.5), numpy (version: 1.19.5), soundfile (version: 0.10.3.post1), opencv (version: 4.5.3.56), moviepy (version: 1.0.3), pydub (version: 0.25.1), matplotlib (version: 3.4.3), time (version: 1.0.0) and tqdm (version: 4.62.2).

Table 4 provides the performance for different combinations of audio and video feature extractor networks based on the classification accuracy. The outcomes show that the combination V8+A4 yields the highest accuracy of 0.86 for the RAVDESS dataset. However, some other well-performing combinations of audio and video feature extractors for this dataset also exist, e.g., V2+A3 (acc = 0.77), V5+A3 (acc = 0.76), V6+A5 (acc = 0.76), and V9+A3 (acc = 0.76), V7+A5 (acc = 0.76), V2+A5 (acc = 0.76), V1+A1 (0.76), V2+A1 (0.75), V4+A1 (0.75), V9+A1 (acc = 0.75), V1+A3 (0.75), V7+A6 (acc = 0.75). The different combinations of audio and video feature extractors have also been experimented with for the SAVEE dataset. The main objective here is to see whether the combination of audio and video feature extractors used in RAVDESS are also well performed, that is, we wanted to see whether there is a generic model or dataset-specific. It is observed that the combination V2+A2 yields the highest accuracy (acc = 0.99) compared to the others. However, similar to RAVDESS, some other well-performing combinations of audio and video feature extractors for SAVEE also exists, e.g., V1+A4 (acc = 0.98), V2+A4 (acc = 0.98), V4+A2 (acc = 0.98), V2+A6 (acc = 0.96), V4+A5 (acc = 0.96), etc.

Detailed performance of the proposed models (i.e., model with V8+A4 for RAVDESS and V2+A2 for SAVEE) along with some other randomly chosen best-performing combinations of audio and video feature extractors are provided in Table 5. The effectiveness of the proposed models is evaluated in terms of several performance measures namely accuracy, specificity, recall, precision, AUC, and F1-score. AUC represents the area under the receiver operating characteristic curve. F1-score indicates the harmonic mean of the macro-averaged recall and precision. The specificity metric, on the other hand, is calculated using Eq. (4).

$$Specificity = \frac{\sum_{k} \frac{tn_k}{tn_k + fp_k}}{\delta} \tag{4}$$

where number of true negatives and false positives for kth emotion class are denoted by tn_k and fp_k respectively. δ represents the number of emotion classes. It is interesting to observe that the model's performance on the SAVEE dataset is relatively high compared to RAVDESS. For instance the best performing model for RAVDESS, i.e, model with combination V8+A4, the values

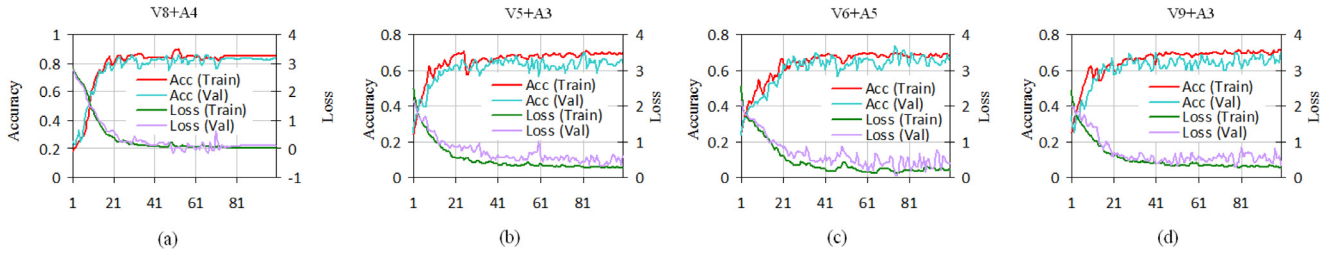

Fig. 4. Accuracy and loss curves for various well-performing combinations of audio and video feature extractor networks for RAVDESS dataset (a) V8+A4 (proposed model) (b) V5+A3 (c) V6+A5 (d) V9+A3.

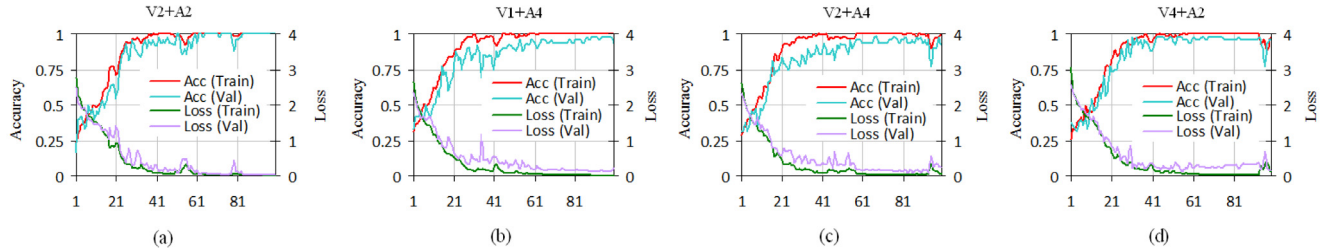

Fig. 5. Accuracy and loss curves for various well-performing combinations of audio and video feature extractor networks for SAVEE dataset (a) V2+A2 (proposed model) (b) V1+A4 (c) V2+A4 (d) V4+A2.

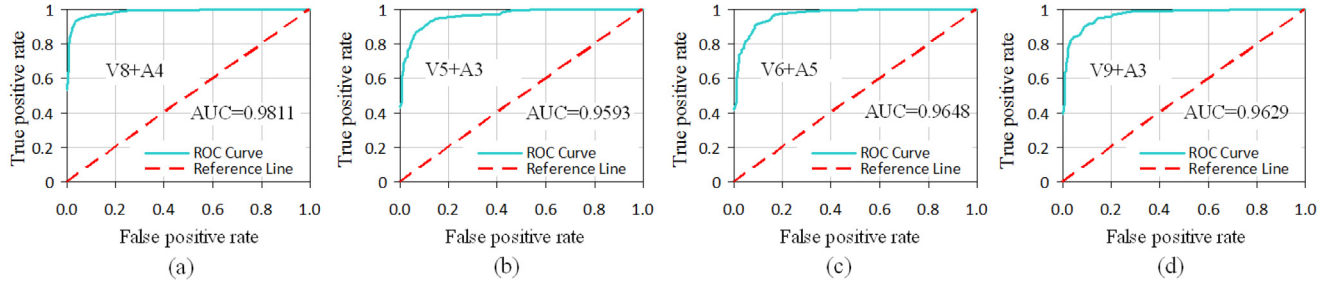

Fig. 6. Receiver operating characteristic (ROC) curves for various well-performing combinations of audio and video feature extractor networks for RAVDESS dataset (a) V8+A4 (proposed model) (b) V5+A3 (c) V6+A5 (d) V9+A3.

of the performance metrics are: accuracy = 0.86, precision = 0.86, recall = 0.86, specificity = 0.98, AUC = 0.98. However, in the case of SAVEE, the performance of the best performing model (accuracy = 0.99, precision = 0.99, recall = 0.99, F1-score = 0.99, specificity = 1, and AUC = 0.99) is much better than that of RAVDESS. This is possible because, unlike RAVDESS, the SAVEE dataset comes with facial landmark data which helps to generate more efficient models. Moreover, it is well known that effective features help in reducing model complexity. Hence, the lightweight V2+A2 model with a total of 8 convolutions and 6 pooling layers was sufficient to achieve the highest accuracy on the SAVEE dataset. On the other hand, as the RAVDESS data had no facial landmarks it was difficult for the face detector to extract relevant facial features from it. Probably that is why the relatively more complex V8+A4 model with a total of 11 convolutions and 6 pooling layers was needed for effective emotion recognition on it.

Now the accuracy (training and validation) and loss (training and validation) curves are generated for various combinations including the proposed models for both the RAVDESS (see Fig. 4) and SAVEE (see Fig. 5) datasets. Note that for both the datasets, the accuracy and loss curves show there are small generalization gaps. In other words, the training and validation curves for both accuracy and loss nearly follow each other. The Receiver Operating Characteristics (ROC) curves for different models for RAVDESS and SAVEE are provided in Figs. 6 and 7 respectively. Plotting true positive rate against false positive rate at different classification thresholds yields a ROC curve. In the ROC plots, the higher the AUC (area under the curve) score, the better a classification model can discriminate between classes. Note that high AUC scores are found for V8+A4 ($\approx$ 0.98) and V2+A2 ($\approx$ 1.00) for RAVDESS and SAVEE respectively. The normalized confusion matrices for different well-performing combinations of audio and video feature extractors are for RAVDESS and SAVEE datasets are presented in Figs. 9 and 10 respectively. The visual inspection quickly confirms that the number of incorrectly classified instances is relatively low.

Moreover, experiments are conducted using different baseline models to investigate the influence of the fusion levels for each dataset. We have performed an extensive exploration of feature-level/model-level and decision-level fusion approaches of four best performing models as well as for the two chosen baselines: (i) 2D CNN-LSTM + 1D CNN and (ii) 2D CNN + 1D CNN. For the first baseline, 2D CNN-LSTM is the video model. On the other hand, for the second baseline, 2D CNN is the video model. For both of them, the 1D CNN is the audio model. The video and audio feature extractors are all vanilla models. Note that the second baseline ignored the temporal aspect of the video features. The method of the feature-level/model-level fusion is as described in Section 3.3 while the decision-level fusion approach is implemented as follows. First, the probability scores of the predictions coming from the video and the audio models are compared. We then choose the emotion class with the higher score as the final predicted emotion label. The experimental results are provided in

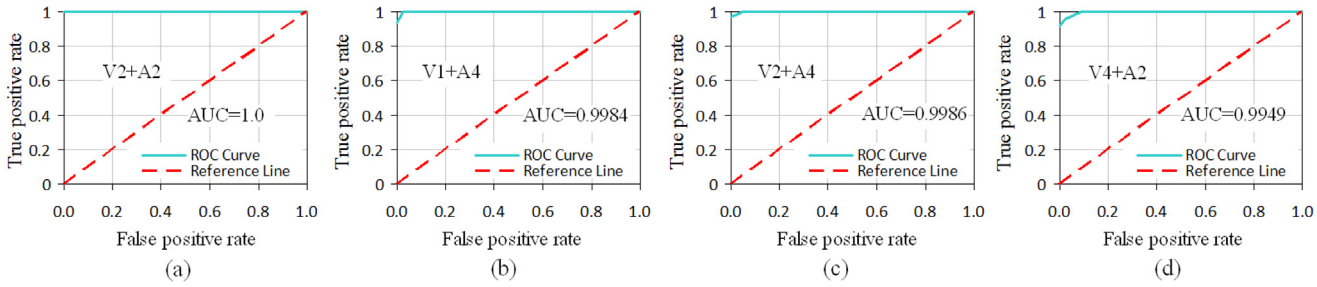

Fig. 7. Receiver operating characteristic (ROC) curves for various well-performing combinations of audio and video feature extractor networks for SAVEE dataset (a) V2+A2 (proposed model) (b) V1+A4 (c) V2+A4 (d) V4+A2.

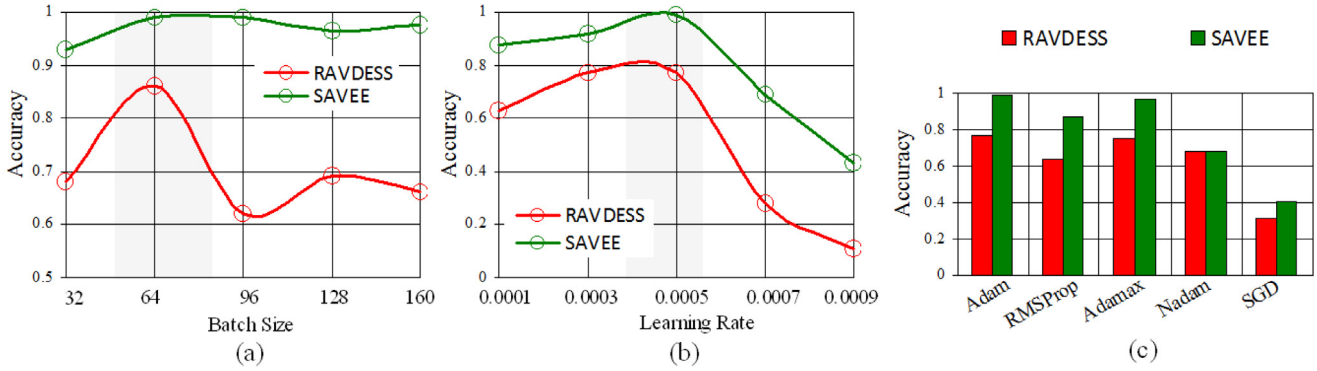

Fig. 8. Impact of model parameters on the proposed models' performance for RAVDESS and SAVEE datasets (a) batch size (b) learning rate and (c) optimizers.

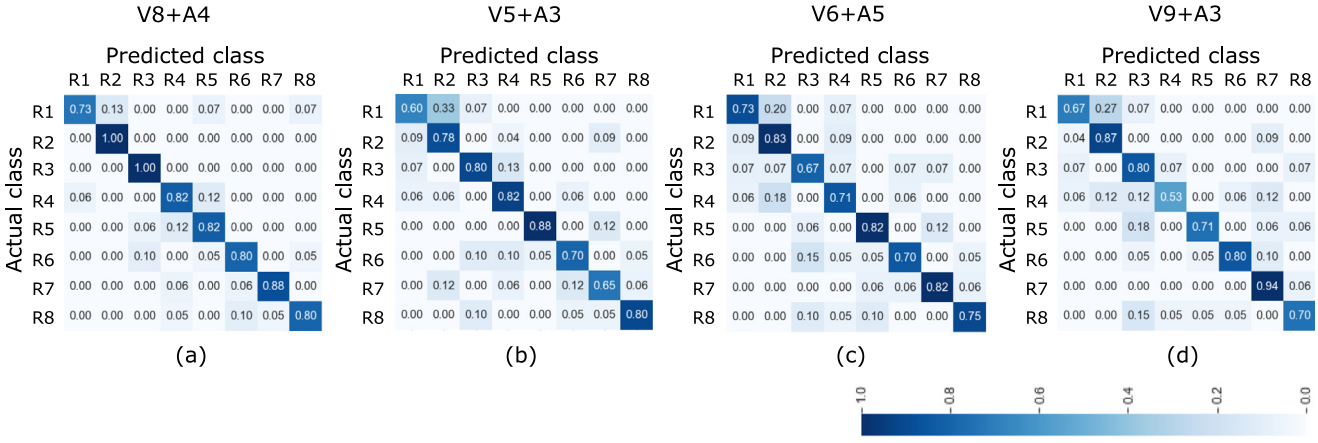

Fig. 9. Normalized confusion matrix for various well-performing combinations of audio and video extractor networks for RAVDESS dataset (a) V8+A4 (proposed model) (b) V5+A3 (c) V6+A5 (d) V9+A3.

Table 6. It is observed that for all the models, the feature/model level fusion approach dominates over the decision level fusion for each dataset. For both the datasets, the proposed models (i.e., V8+A4 for RAVDESS and V2+A2 for SAVEE) achieve the best performance for both the fusion approaches.

4.1. Impact of model parameters

This section explores how model parameters influence the effectiveness of the proposed techniques. In other words, sensitivity analysis of the proposed models is conducted. The experiment is carried out based on three model parameters namely batch size, learning rate, and optimizer. The model's performance is found to be significantly influenced by all three factors. The results of the experiments for batch size, learning rate, and optimizer on the RAVDESS and SAVEE models are provided in Fig. 8(a), (b), and (c) respectively. As given in Fig. 8(a), batch sizes of 32, 64, 96, 128, and 160 are used to compute the models' performance. It is observed that if the batch size is 64, the highest predictive accuracy is achieved for both datasets. It is worth noting that as the batch size grows larger (i.e., batch size $>$ 64), the accuracy decreases. In this context, existing literature [60] also highlights the fact that increasing the batch size does not always guarantee high predictive accuracy rather it may result in its degradation. Depending on various factors like the complexity of the dataset involved (noisy, homogeneous, heterogeneous, etc.), learning rate and optimizer the accuracy may rise or fall with increasing batch size. Learning rate of 0.0001, 0.0003, 0.0005, 0.0007, and 0.0009 are used to assess the model's performance (see Fig. 8(b)). The proposed models are also evaluated with different optimizers (see Fig. 8(c)), such as SGD, Nadam, Adamax, RMSProp, and Adam. For both datasets, the proposed models are providing better results for the Adam optimizer.

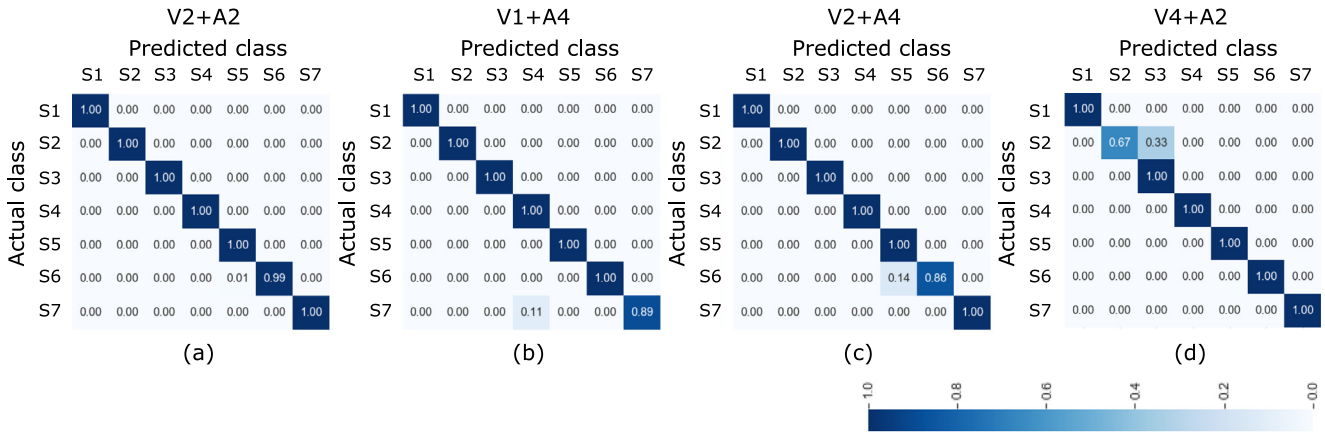

Fig. 10. Normalized confusion matrix for various well-performing combinations of audio and video extractor networks for SAVEE dataset (a) V2+A2 (proposed model) (b) V1+A4 (c) V2+A4 (d) V4+A2.

Table 6
Accuracy of the models on fusion levels for each dataset.

Dataset	Model	Fusion level	Accuracy
RAVDESS	V8 + A4 (proposed)	Feature/model-level	0.86
		Decision-level	0.79
	V5 + A3	Feature/model-level	0.76
		Decision-level	0.73
	V6 + A5	Feature/model-level	0.76
		Decision-level	0.66
	V9 + A3	Feature/model-level	0.76
		Decision-level	0.71
	2D CNN-LSTM + 1D CNN	Feature/model-level	0.77
		Decision-level	0.73
	2D CNN + 1D CNN	Feature/model-level	0.78
		Decision-level	0.71
SAVEE	V2 + A2 (proposed)	Feature/model-level	0.99
		Decision-level	0.97
	V1 + A4	Feature/model-level	0.98
		Decision-level	0.91
	V2 + A4	Feature/model-level	0.98
		Decision-level	0.97
	V4 + A2	Feature/model-level	0.98
		Decision-level	0.93
	2D CNN-LSTM + 1D CNN	Feature/model-level	0.93
		Decision-level	0.83
	2D CNN + 1D CNN	Feature/model-level	0.89
		Decision-level	0.83

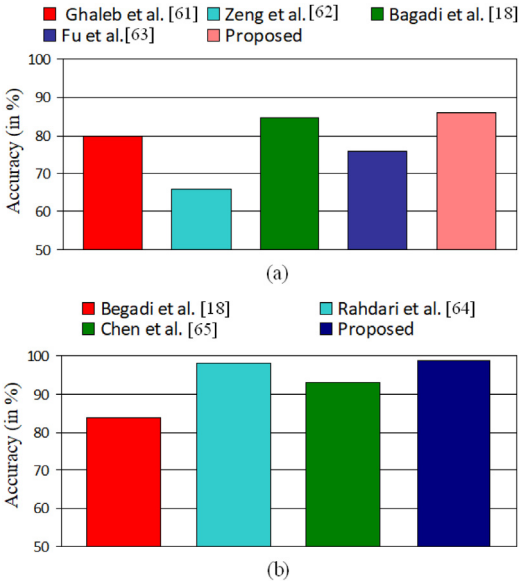

Fig. 11. Performance comparison of proposed models with the previous works for (a) RAVDEES (b) SAVEE.

4.2. Performance comparison with the existing works

The accuracy of previous approaches is compared to the proposed models in Fig. 11(a) for the RAVDESS dataset. Ghaleb et al. [61] introduced a novel multimodal model based on DML (Deep Metric Learning) for effective multimodal emotion recognition, achieving the highest accuracy of 80%. Zeng et al. [62] fed spectrograms to their proposed deep neural network (DNN) based emotion recognition model An accuracy score of 65.97% is achieved on the validation set. Bagadi et al. [18] employed an LSTM for multimodal emotion classification from the audio and textual information, achieving an accuracy of 85.34%. Fu et al. [63] have performed a multimodal emotion detection using a cross-modal fusion network with a total of 26.30 million parameters. They attained an accuracy of 75.76% on the RAVDESS dataset. Our proposed work uses audio–visual information from the 1440 audio–visual files of the RAVDESS dataset and achieves a test accuracy of 86.11%. Moreover, our proposed model has only 6,028,104 parameters, which is significantly less than previous work. Hence, the proposed model is more efficient and computationally light and well suited for real-time applications.

The accuracy of previous approaches is compared to the proposed models in Fig. 11(b) for the SAVEE dataset. Bagadi et al. [18] also employed the LSTM for multimodal emotion classification from audio–textual information for the SAVEE dataset, achieving an accuracy of 83.7%. Rahdari et al. [64] utilized facial landmark features as the visual features and prosodic and spectral features as the audio features for their work. They concatenated all these features and fed them to different classifiers to find out the optimal multimodal model. Their best model was an FRNN (Fuzzy Rough Neural Network) achieving an accuracy of 98%. Chen et al. [65] extracted the audio–visual features and fused them using a novel K-means clustering + KCCA (Kernel Canonical Correlation Analysis) method. Their approach achieved an accuracy of 93.06%. Our method outperforms all previous works on this dataset using the 480 total audio–visual files, achieving the test accuracy of 99%.

Several factors contributed to the proposed model's ability to outperform the previous works in this domain. For instance,

Table 7
The details of the publicly available audio–visual media for the case study.

Id	Description	Source	
		Original video	Extracted video
VD_1	A speech by Barack Obama, former US President	https://bit.ly/3mcvS4v	https://bit.ly/39TJrQI
VD_2	A speech by Sourav Ganguly, former captain of Indian cricket team	https://bit.ly/3uxFVFm	https://bit.ly/3ioRsln
VD_3	A speech by Netaji Subhash Chandra Bose, the valiant freedom fighter of India	https://bit.ly/39PAXu1	https://bit.ly/2Yir9X1
VD_4	A speech by Professor Angus Deaton, a renowned economist	https://bit.ly/2XXXB0t	https://bit.ly/3D2feLR

Ghaleb et al. [61] and Bagadi et al. [18] employ LSTM models for capturing the temporal aspect of the video features. In contrast to this, we use a ConvLSTM2D layer in our video feature extractor model to effectively deal with the video features. This layer is basically a merger of 2-dimensional convolution layer and LSTM layer. The benefit of this layer is that it can leverage the advantages of both an LSTM and a CNN and hence is more robust than the individual models. Hence this might be the reason why we were able to capture the visual as well as temporal aspects of our video features more effectively than [18, 61]. Zeng et al. [62] use only melspectrograms and Fu et al. [63] use only MFCC features as the audio features in their work. On the contrary, we extracted several important audio features like MFCC, melspectrogram, spectral contrast and tonnetz that we fed to our multimodal model. Thus having more effective audio features might have impacted the performance of our model. Finally, Rahdari et al. [64] and Chen et al. [65] respectively employ facial landmark analysis and Viola–Jones algorithm for facial feature extraction. On the other hand, we used a more accurate and sophisticated inbuilt deep learning based face detector of the computer vision library OpenCV for efficient visual feature extraction. Due to the above-mentioned reasons, our proposed multimodal model was able to recognize human emotions with a higher accuracy.

4.3. Case studies

Some experiments are performed on publicly available audio–visual media to capture the variability in the emotional states. In other words, the main aim here is to observe whether the models can detect the changes in the emotional states over time for completely new videos on which they have not been trained. All our videos for this case study have been taken from YouTube. The proposed models have been specifically trained to detect emotion for a single person over time, hence we had to trim the videos to leave out the unnecessary faces and merge them accordingly so that we get the face of a single speaker throughout the entire length of the video. For this purpose, the video samples of 90 s length are extracted from all the original ones. The sample video length is kept the same for each extracted video to maintain uniformity. The details of each video are given in Table 7. A brief description, original source, and source of the extracted video on which the case study is conducted are provided for each video.

Fig. 12 shows the overall dominance of various emotional states over the 90 s of the video contents. The emotional dominance (E_d^k) of kth emotional state for any video is computed based on the following expression.

$$E_d^k = \frac{1}{N}\sum_{i=1}^{N} p_i^k \tag{5}$$

Here, the number N denotes how many predictions have been made for each video. Note that our models are trained to predict the emotions detected for every 6 frames of any video. As each of the videos is 90 s long we get a total of 15 frames in batches of 6 and thus 15 predictions for each video are there. The term p_i^k denotes the softmax probability score of kth emotional states for ith prediction. As shown in Fig. 12, for VD_1 (i.e., Barack Obama), VD_2 (i.e., Sourav Ganguly), VD_3 (i.e., Netaji Subhash Chandra Bose), and VD_4 (Angus Deaton) the dominating emotional states are disgust, neutral, angry, and calm respectively. It is worth noting that since these publicly available video data do not come with facial landmark data, it will not be suitable to predict using the model which is based on the SAVEE dataset. Hence, the case study is performed based on the proposed model of RAVDESS with V8+A4 combination of feature extractors. Now, in Fig. 13, the variation in the emotional states of the speakers over the 90-second time interval for various videos is shown.

5. Conclusion

This work investigates the model-based fusion approach of audio–visual modalities in order to determine the best multimodal model for classifying human emotions from audio and video data. Various combinations of audio and video feature extractor networks are analyzed as part of this process. The efficacy of the proposed models is validated by comparing their performance with the previous works in this domain. Note that two different multimodal models achieve the best results on the two datasets, RAVDESS (1440 audio–visual files) and SAVEE (480 audio–visual files), proving that finding a universally optimal model for multimodal emotion recognition is quite challenging. These challenges make this domain fascinating and hence it still remains the subject of intensive research.

The spectral audio features are mainly considered in this work, which may be a limitation as there are more auditory features that could have been utilized. In the future, we will aim to address this limitation. We would also like to perform multimodal emotion recognition using text data along with audio–visual data as our future work.

CRediT authorship contribution statement

Asif Iqbal Middya: Conceptualization, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Baibhav Nag:** Methodology, Software, Validation, Writing – original draft. **Sarbani Roy:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The research work of Asif Iqbal Middya is supported by UGC-NET Junior Research Fellowship (UGC-Ref. No.: 3684/(NET-JULY 2018)) provided by the University Grants Commission, Government of India.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

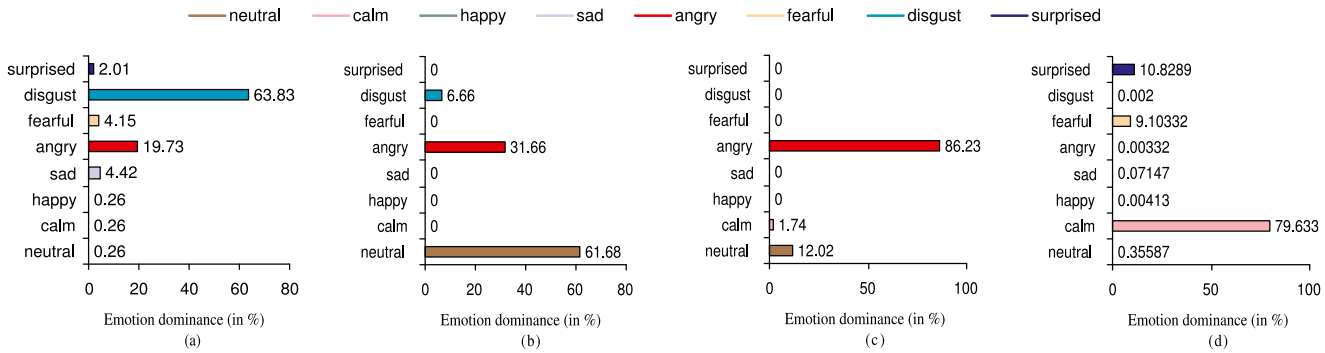

Fig. 12. Emotional dominance of the emotional states over time for different videos (a) VD_1 i.e. Barack Obama (b) VD_2 i.e., Sourav Ganguly (c) VD_3 i.e., Netaji Subhash Chandra Bose and (d) VD_4 i.e., Angus Deaton.

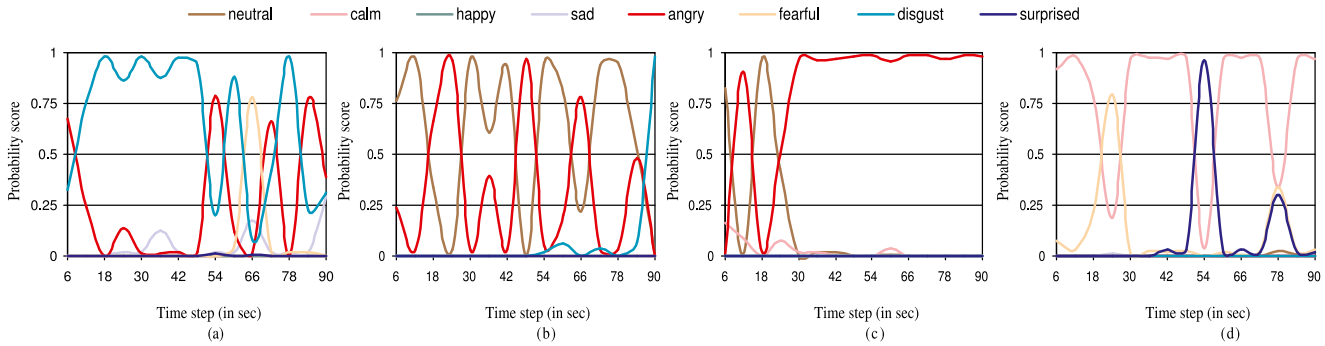

Fig. 13. Variability of the emotional states over time for different videos (a) VD_1 i.e. Barack Obama (b) VD_2 i.e., Sourav Ganguly (c) VD_3 i.e., Netaji Subhash Chandra Bose and (d) VD_4 i.e., Angus Deaton.

References

[1] M.S. Hossain, G. Muhammad, Emotion recognition using deep learning approach from audio–visual emotional big data, Inf. Fusion 49 (2019) 69–78.

[2] R.W. Picard, Affective Computing, MIT Press, 2000.

[3] A.A. Varghese, J.P. Cherian, J.J. Kizhakkethottam, Overview on emotion recognition system, in: 2015 International Conference on Soft-Computing and Networks Security (ICSNS), IEEE, 2015, pp. 1–5.

[4] K.R. Scherer, et al., Psychological models of emotion, Neuropsychol. Emot. 137 (3) (2000) 137–162.

[5] Q. Li, Z. Yang, S. Liu, Z. Dai, Y. Liu, The study of emotion recognition from physiological signals, in: 2015 Seventh International Conference on Advanced Computational Intelligence (ICACI), IEEE, 2015, pp. 378–382.

[6] X. Huang, J. Kortelainen, G. Zhao, X. Li, A. Moilanen, T. Seppänen, M. Pietikäinen, Multi-modal emotion analysis from facial expressions and electroencephalogram, Comput. Vis. Image Underst. 147 (2016) 114–124.

[7] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, M. Pantic, Avec 2016: Depression, mood, and emotion recognition workshop and challenge, in: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, 2016, pp. 3–10.

[8] M. Mukeshimana, X. Ban, N. Karani, R. Liu, Multimodal emotion recognition for human-computer interaction: A survey, System 9 (2017) 10.

[9] E. Avots, T. Sapiński, M. Bachmann, D. Kamińska, Audiovisual emotion recognition in wild, Mach. Vis. Appl. 30 (5) (2019) 975–985.

[10] S.-u. Haq, Audio Visual Expressed Emotion Classification, University of Surrey, (United Kingdom), 2011.

[11] S.R. Livingstone, F.A. Russo, 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) (2018) e0196391.

[12] A. Ortony, G.L. Clore, A. Collins, The Cognitive Structure of Emotions, Cambridge University Press, 1990.

[13] P. Shaver, J. Schwartz, D. Kirson, C. O'connor, Emotion knowledge: further exploration of a prototype approach., J. Personal. Soc. Psychol. 52 (6) (1987) 1061.

[14] E. Cambria, A. Livingstone, A. Hussain, The hourglass of emotions, in: Cognitive Behavioural Systems, Springer, 2012, pp. 144–157.

[15] P. Ekman, An argument for basic emotions, Cogn. Emot. 6 (3–4) (1992) 169–200.

[16] A.T. Latinjak, The underlying structure of emotions: A tri-dimensional model of core affect and emotion concepts for sports, Rev. Iberoam. Psicol. Ejerc. Deporte 7 (1) (2012) 71–88.

[17] R. Plutchik, The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice, Am. Sci. 89 (4) (2001) 344–350.

[18] K.R. Bagadi, A comprehensive analysis of multimodal speech emotion recognition, in: Journal of Physics: Conference Series, Vol. 1917, IOP Publishing, 2021, 012009.

[19] M. Abdullah, M. Ahmad, D. Han, Facial expression recognition in videos: An CNN-LSTM based model for video classification, in: 2020 International Conference on Electronics, Information, and Communication (ICEIC), IEEE, 2020, pp. 1–3.

[20] D. Krishna, A. Patil, Multimodal emotion recognition using cross-modal attention and 1d convolutional neural networks, in: Interspeech, 2020, pp. 4243–4247.

[21] A. Jaratrotkamjorn, A. Choksuriwong, Bimodal emotion recognition using deep belief network, in: 2019 23rd International Computer Science and Engineering Conference (ICSEC), IEEE, 2019, pp. 103–109.

[22] G. Sahu, Multimodal speech emotion recognition and ambiguity resolution, 2019, arXiv preprint arXiv:1904.06022.

[23] K.P. Rao, M.C.S. Rao, N.H. Chowdary, An integrated approach to emotion recognition and gender classification, J. Vis. Commun. Image Represent. 60 (2019) 339–345.

[24] S. Yoon, S. Byun, K. Jung, Multimodal speech emotion recognition using audio and text, in: 2018 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2018, pp. 112–118.

[25] D. Nguyen, K. Nguyen, S. Sridharan, D. Dean, C. Fookes, Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition, Comput. Vis. Image Underst. 174 (2018) 33–42.

[26] H. Miao, Y. Zhang, W. Li, H. Zhang, D. Wang, S. Feng, Chinese multimodal emotion recognition in deep and traditional machine leaming approaches, in: 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), IEEE, 2018, pp. 1–6.

[27] F. Xu, Z. Wang, Emotion recognition research based on integration of facial expression and voice, in: 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), IEEE, 2018, pp. 1–6.

[28] J. Yan, W. Zheng, Q. Xu, G. Lu, H. Li, B. Wang, Sparse kernel reduced-rank regression for bimodal emotion recognition from facial expression and speech, IEEE Trans. Multimed. 18 (7) (2016) 1319–1329.

[29] S.E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R.C. Ferrari, et al., Combining modality specific deep neural networks for emotion recognition in video, in: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, 2013, pp. 543–550.
[30] N. Srivastava, R. Salakhutdinov, et al., Multimodal learning with deep Boltzmann machines, in: NIPS, Vol. 1, Citeseer, 2012, p. 2.
[31] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Ng, Multimodal deep learning, in: International Conference on Machine Learning (ICML), Bellevue, WA, 2011, pp. 689–696.
[32] Y. Wang, L. Guan, Recognizing human emotional state from audiovisual signals, IEEE Trans. Multimed. 10 (5) (2008) 936–946.
[33] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, S. Narayanan, Analysis of emotion recognition using facial expressions, speech and multimodal information, in: Proceedings of the 6th International Conference on Multimodal Interfaces, 2004, pp. 205–211.
[34] Y. Yoshitomi, S.-I. Kim, T. Kawano, T. Kilazoe, Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face, in: Proceedings 9th IEEE International Workshop on Robot and Human Interactive Communication. IEEE RO-MAN 2000 (Cat. No. 00TH8499), IEEE, 2000, pp. 178–183.
[35] Z. Wang, S.-B. Ho, E. Cambria, A review of emotion sensing: categorization models and algorithms, Multimedia Tools Appl. 79 (47–48) (2020) 35553–35582, http://dx.doi.org/10.1007/s11042-019-08328-z.
[36] A. Ortony, T.J. Turner, What's basic about basic emotions? Psychol. Rev. 97 (3) (1990) 315.
[37] B.R. Steunebrink, M. Dastani, J.J.C. Meyer, The OCC model revisited, in: Proceedings of the 4th Workshop on Emotion and Computing, 2009.
[38] Y. Li, J. Tao, L. Chao, W. Bao, Y. Liu, CHEAVD: a Chinese natural emotional audio–visual database, J. Ambient Intell. Humaniz. Comput. 8 (6) (2017) 913–924.
[39] O. Martin, I. Kotsia, B. Macq, I. Pitas, The eNTERFACE' 05 audio-visual emotion database, in: 22nd International Conference on Data Engineering Workshops (ICDEW'06), 2006, p. 8, http://dx.doi.org/10.1109/ICDEW.2006.145.
[40] E. Patterson, S. Gurbuz, Z. Tufekci, J. Gowdy, CUAVE: A new audio-visual database for multimodal human-computer interface research, in: 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, 2002, pp. II–2017–II–2020, http://dx.doi.org/10.1109/ICASSP.2002.5745028.
[41] I. Matthews, T.F. Cootes, J.A. Bangham, S. Cox, R. Harvey, Extraction of visual features for lipreading, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2) (2002) 198–213.
[42] A. Dhall, R. Goecke, J. Joshi, M. Wagner, T. Gedeon, Emotion recognition in the wild challenge 2013, in: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, 2013, pp. 509–516.
[43] J. Huang, J. Tao, B. Liu, Z. Lian, M. Niu, Multimodal transformer fusion for continuous emotion recognition, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, http://dx.doi.org/10.1109/icassp40776.2020.9053762.
[44] J.-B. Delbrouck, N. Tits, M. Brousmiche, S. Dupont, A transformer-based joint-encoding for emotion recognition and sentiment analysis, 2020, arXiv preprint arXiv:2006.15955.
[45] Y. Susanto, A.G. Livingstone, B.C. Ng, E. Cambria, The hourglass model revisited, IEEE Intell. Syst. 35 (5) (2020) 96–102.
[46] M. Dragoni, S. Poria, E. Cambria, OntoSenticNet: A commonsense ontology for sentiment analysis, IEEE Intell. Syst. 33 (3) (2018) 77–85.
[47] M. Dragoni, I. Donadello, E. Cambria, Ontosenticnet 2: enhancing reasoning within sentiment analysis, IEEE Intell. Syst. 36 (2021) 5.
[48] E. Cambria, N. Howard, J. Hsu, A. Hussain, Sentic blending: Scalable multimodal fusion for the continuous interpretation of semantics and sentics, in: 2013 IEEE Symposium on Computational Intelligence for Human-Like Intelligence (CIHLI), IEEE, 2013, http://dx.doi.org/10.1109/cihli.2013.6613272.
[49] I. Chaturvedi, R. Satapathy, S. Cavallari, E. Cambria, Fuzzy commonsense reasoning for multimodal sentiment analysis, Pattern Recognit. Lett. 125 (2019) 264–270, http://dx.doi.org/10.1016/j.patrec.2019.04.024.
[50] L. Stappen, A. Baird, E. Cambria, B.W. Schuller, E. Cambria, Sentiment analysis and topic recognition in video transcriptions, IEEE Intell. Syst. 36 (2) (2021) 88–95, http://dx.doi.org/10.1109/mis.2021.3062200.
[51] K. Zhang, Y. Li, J. Wang, E. Cambria, X. Li, Real-time video emotion recognition based on reinforcement learning and domain knowledge, IEEE Trans. Circuits Syst. Video Technol. (2021) 1, http://dx.doi.org/10.1109/tcsvt.2021.3072412.
[52] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[53] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017.
[54] OpenCV, Open Source Computer Vision Library, 2015.
[55] J. Robert, M. Webbie, et al., Pydub, 2018, URL http://pydub.com/.
[56] S.S. Stevens, J. Volkmann, E.B. Newman, A scale for the measurement of the psychological magnitude pitch, J. Acoust. Soc. Am. 8 (3) (1937) 185–190.
[57] Y. Soon, S.N. Koh, C.K. Yeo, Noisy speech enhancement using discrete cosine transform, Speech Commun. 24 (3) (1998) 249–257.
[58] D.-N. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, L.-H. Cai, Music type classification by spectral contrast feature, in: Proceedings. IEEE International Conference on Multimedia and Expo, Vol. 1, IEEE, 2002, pp. 113–116.
[59] Y. Li, J. Yang, J. Wen, Entropy-based redundancy analysis and information screening, Digit. Commun. Netw. (2021) http://dx.doi.org/10.1016/j.dcan.2021.12.001.
[60] I. Kandel, M. Castelli, The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset, ICT Express 6 (4) (2020) 312–315.
[61] E. Ghaleb, M. Popa, S. Asteriadis, Multimodal and temporal perception of audio-visual cues for emotion recognition, in: 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2019, pp. 552–558.
[62] Y. Zeng, H. Mao, D. Peng, Z. Yi, Spectrogram based multi-task audio classification, Multimedia Tools Appl. 78 (3) (2019) 3705–3722.
[63] Z. Fu, F. Liu, H. Wang, J. Qi, X. Fu, A. Zhou, Z. Li, A cross-modal fusion network based on self-attention and residual structure for multimodal emotion recognition, 2021, arXiv:2111.02172.
[64] F. Rahdari, E. Rashedi, M. Eftekhari, A multimodal emotion recognition system using facial landmark analysis, Iran. J. Sci. Technol. Trans. Electr. Eng. 43 (1) (2019) 171–189.
[65] L. Chen, K. Wang, M. Wu, W. Pedrycz, K. Hirota, K-means clustering-based kernel canonical correlation analysis for multimodal emotion recognition, IFAC-PapersOnLine 53 (2) (2020) 10250–10254.