arXiv:2208.11868v2 [cs.CV] 7 Jan 2023

Interpretable Multimodal Emotion Recognition using Hybrid Fusion of Speech and Image Data

Puneet Kumar[1*†], Sarthak Malik[2†] and Balasubramanian Raman[1]

[1]Department of Computer Science and Engineering, Indian Institute of Technology Roorkee.
[2]Department of Electrical Engineering, Indian Institute of Technology Roorkee.

*Corresponding author(s). E-mail(s): pkumar99@cs.iitr.ac.in;
Contributing authors: sarthak_m@mt.iitr.ac.in; bala@cs.iitr.ac.in;
[†]These authors contributed equally to this work.

Abstract

This paper proposes a multimodal emotion recognition system based on hybrid fusion that classifies the emotions depicted by speech utterances and corresponding images into discrete classes. A new interpretability technique has been developed to identify the important speech & image features leading to the prediction of particular emotion classes. The proposed system's architecture has been determined through intensive ablation studies. It fuses the speech & image features and then combines speech, image, and intermediate fusion outputs. The proposed interpretability technique incorporates the divide & conquer approach to compute shapely values denoting each speech & image feature's importance. We have also constructed a large-scale dataset (IIT-R SIER dataset) consisting of speech utterances, corresponding images, and class labels, i.e., 'anger,' 'happy,' 'hate,' and 'sad.' The proposed system has achieved 83.29% accuracy for emotion recognition. The enhanced performance of the proposed system advocates the importance of utilizing complementary information from multiple modalities for emotion recognition.

Keywords: Affective Computing, Multimodal Analysis, Speech and Image Processing, Interpretable AI, Information Fusion.

1 Introduction

The multimedia data has overgrown in the last few years, leading multimodal emotion analysis to emerging as an important research trend [1]. The need to develop multimodal emotion processing systems capable of recognizing various emotions from images and texts is rapidly increasing. Research in this direction aims to help machines become empathetic as emotion analysis is used in various applications such as cognitive psychology, automated identification, intelligent devices, and human-machine interface [27]. The speech and image modalities portray human emotions and intentions very effectively [12]. Combining complementary information from both of these modalities could increase emotion recognition accuracy [42].

Researchers have attempted to identify emotions by processing audio and visual information separately [18, 28, 39]. However, multimodal emotion recognition, where the emotional context from multiple modalities are analyzed together, performs better than unimodal emotion recognition [42]. In this context, multimodal emotion recognition from speech & text modalities and image & text modalities have been performed; however, emotion recognition from speech & image modalities has yet to be fully explored. Moreover, most of the existing multimodal approaches do not focus on interpreting the internal working of their emotion recognition systems. It inspired us to develop a multimodal emotion recognition system capable of recognizing emotions portrayed by speech utterances & corresponding images and explaining the importance of each speech segment & visual feature towards emotion recognition.

Multimodal emotion recognition also faces the issue of the unavailability of sufficient labeled datasets for training. Moreover, the real-life multimodal data contains generic images with facial, human, and non-human objects, but most of the existing multimodal datasets contain only facial and human images [2]. A few multimodal datasets are available that contain generic images; however, they consist of positive, negative, and neutral sentiment labels and do not contain multi-class emotion labels [5, 37]. A new dataset, 'IIT Roorkee Speech & Image Emotion Recognition (IIT-R SIER) dataset,' has been constructed to address this issue. It contains generic images, corresponding speech utterances, and discrete class labels, i.e., 'anger,' 'happy,' 'hate,' and 'sad.' We used the data instances with identical predicted emotion labels for image and text modalities to construct the dataset. This paper analyses the improvements in SER on combining the complementary information from corresponding images.

The proposed system, 'ParallelNet,' recognizes emotions in speech utterances and corresponding images. It implements two networks, $N1$ and $N2$, to fuse the information of speech and image modalities in a hybrid manner of intermediate and late fusion. The architectures for $N1$ and $N2$ are determined through extensive ablation studies. A technique to interpret the important input features and predictions has also been developed. The proposed system has performed with an accuracy of 83.29% on the IIT-R SIER dataset. The dataset and code for this paper are accessible at https://github.com/MIntelligence-Group/SpeechImg_EmoRec.

The paper makes the following major contributions.

- A hybrid-fusion-based novel system, 'ParallelNet,' has been proposed to classify an input containing speech utterance & corresponding image into discrete emotion classes. It combines the information from speech & image modalities using a hybrid of intermediate and late fusion.

- A large-scale dataset, 'IIT-R SIER dataset' containing speech utterances, corresponding images, and emotion labels, has been constructed.

- A new interpretability technique has been developed to identify the important parts of the input speech and image that contribute the most to recognizing emotions.

Further in this paper, the related works have been reviewed in Section 2. The proposed dataset, system, and interpretability technique have been described in Section 3 along with the dataset compilation procedure. Section 4 and 5 discuss the experiments and results. Finally, Section 6 concludes the paper and highlights the directions for future research.

2 Related works

This Section surveys the existing literature on speech & image emotion recognition and the interpretability of deep neural networks.

2.1 Speech emotion recognition

The deep learning-based approaches using spectrogram features and attention mechanisms have shown state-of-the-art results for speech emotion recognition (SER) [4, 13, 40]. In this context, Xu et al. [39] generated multiple attention maps, fused and used them for SER. They observed an increased performance as compared to non-fusion-based approaches. In another work, Majumder et al. [18] implemented a deep neural network to track speakers' identities showing specific emotions.

2.2 Image emotion recognition

Image Emotion Recognition (IER) research is also an active domain. For instance, Kim et al. [12] built a deep feed-forward neural network to combine different levels of emotion features obtained by using the semantic information of the image. In another work, Rao et al. [28] prepared hierarchical notations for emotion recognition in the visual domain.

The human emotions can be expressed in various modalities, out of which speech & image express the emotional intentions most effectively [12]. Analysis in a single modality may not be able to recognize the emotional context completely, which leads to the need for multimodal emotion recognition approaches that analyze multimodal audio-visual emotional context [42].

2.3 Multimodal emotion recognition

Multimodal emotion analysis from audio-visual data has started getting researchers' attention lately [6, 8, 15]. For instance, Siriwardhana et al. [34] fine-tuned Transformers-based models to improve the performance of multimodal speech emotion recognition. Multimodal emotion recognition has been carried out for text & speech modalities [14, 19] and text & image modalities [5, 15, 37]. However, it has not been fully explored for speech & image modalities. Moreover, most deep learning-based multimodal emotion recognition systems work as a black box where it is difficult to interpret their inside mechanism. It inspired us to develop an interpretable multimodal emotion recognition system for speech & image modalities.

2.4 Interpretability of deep neural networks

The existing interpretability approaches compute each input feature's importance by backpropagating the network or observing the changes in output on changing the input [17]. In this direction, Riberio et al. [29] explained a network based on each input's importance. Researchers have explained the layer-by-layer learning of deep neural networks and the output based on all the neurons' contributions [14, 32]. There are interpretability methods for visual analysis to compute input pixels' importance [17, 20, 29]. However, such methods still need to be sufficiently explored for speech modality. It inspired us to develop an interpretability technique for multimodal emotion recognition to explain the importance of each speech segment and each visual feature of the input.

3 Proposed Methodology

3.1 Dataset construction

The 'IIT Roorkee Speech & Image Emotion Recognition (IIT-R SIER) dataset has been constructed using Balanced Twitter for Sentiment Analysis (B-T4SA) dataset [37]. The recent text-to-speech models generate high-quality audio that can be used as a valid approximation of natural audio signals [21, 22, 25]. A pre-trained state-of-the-art text-to-speech model, DeepSpeech3 [25], has been used to convert the text from the B-T4SA dataset to speech. The samples are manually cleaned by removing the corrupt and duplicate samples. Further, the following procedure has been followed to generate the ground-truth labels according to the overall emotional context represented by both modalities in combination.

Various parameters of the SIER dataset have been summarised in Fig. 1 whereas the procedure to construct the same has been described as follows.

The speech component of each data sample is passed through the SER model trained on The Interactive Emotional Dyadic Motion Capture (IEMOCAP) [2] dataset; classification probabilities for each emotion class are obtained, and the maximum among the probabilities for all emotion classes is noted as max_1. Likewise, each sample's image component is passed through the IER model trained on Flickr & Instagram (FI) [41] dataset, and the maximum of classification probabilities for all
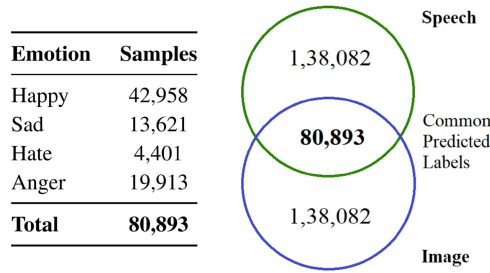
Emotion	Samples
Happy	42,958
Sad	13,621
Hate	4,401
Anger	19,913
Total	**80,893**

Fig. 1: Summary of IIT-R SIER dataset. Left: Class-wise data samples distribution. Right: Modality-wise data distribution.

emotion classes, i.e., max_2 is noted. The higher of max_1 and max_2 is observed, and the corresponding emotion label is assigned as the ground-truth label to the data sample. For example, if IER model returned probabilities 0.1, 0.8, 0.05 and 0.05 for four emotion classes while SER model gave 0.1, 0.1, 0.7 & 0.1 then we assigned second emotion class to the sample considering $max(0.8$ and $0.7)$. The samples having $max(max_1, max_2)$ less than a threshold of 0.5 are discarded as the predicted class label must be at least double confident than random prediction (probability 0.25). The samples labeled as 'excitement' & 'disgust' have been re-labeled as 'happy' & 'hate' as per Plutchik's wheel of emotions [26]. The final dataset contains a total of 80,893 samples with 42,958 labeled as 'happy,' 13621 as 'sad', and 4401 & 19,913 as 'hate' and 'anger' respectively.

We did not take the samples with the same predicted labels by SER and IER, as speech & image modalities might favor different emotion classes in isolation. In contrast, we are interested in the emotion class denoted by both modalities together. Samples for which SER and IER models predicted the same emotion label have been retained to form the IIT-R SIER dataset. The samples having the same predicted labels for SER and IER models denote high confidence in both modalities. They have been retained irrespective of whether the labels are correct. This approach is inspired by the B-T4SA dataset's base paper, where the samples having high confidence in text and image modalities are kept while others are discarded [37].

3.1.1 Human evaluation

We had two human readers (one male and one female) who spoke out and recorded the text components of the data samples. The evaluators listened to the machine-synthesized and human speech recorded by the human readers and labeled the emotion classes portrayed by them. The samples have been picked randomly, and the average of the evaluators' scores has been reported in Table 1. Here, A_i denotes the emotion classification accuracy when the human evaluators predicted the emotions considering the image components. Likewise, A_{ss} & A_{hs} are the accuracy values on considering the synthetic and human speech components, and A_{ss-i} & A_{hs-i} are the accuracies on considering both speech and image modalities.

Table 1: Human evaluation of SIER dataset. Where A_m denotes the emotion classification accuracy for modality m, i: image modality, ss: synthetic speech, hs: human speech, $ss - i$: multimodal context combining synthetic speech and image modalities and $hs - i$: multimodal context combining human speech and image modalities.

Class	A_i	A_{ss}	A_{hs}	A_{ss-i}	A_{hs-i}
Happy	63.89%	66.67%	69.44%	73.48%	75.87%
Sad	75.00%	77.08%	78.13%	82.43%	83.27%
Hate	67.86%	71.43%	72.32%	77.64%	81.32%
Anger	70.31%	82.81%	85.94%	82.17%	84.19%
Overall	69.26%	74.49%	76.46%	78.93%	80.46%

The following two major observations can be drawn from Table 1: **i)** The similar values of 74.49% for synthetic speech and 78.91% for human speech advocate that the speech component of the data generated through text-to-speech is mature enough and embodies the appropriate emotional context. **ii)** Considering complementary information from both speech and image modalities led to higher emotion recognition performance. The evaluators also reported that 78.93% of the samples considering machine-synthesized speech along with the corresponding image were in line with the determined emotion label, whereas this is comparable to the value of 80.46% on considering human speech along with the corresponding image with is significantly higher than the accuracies on considering only image or only speech components.

3.2 Proposed multimodal emotion recognition system

Fig. 2 depicts the architecture of the proposed multimodal emotion recognition system, which is determined in Section 4.3 through the ablation studies. A hybrid of intermediate and late fusion is implemented where intermediate fusion combines various modalities' information before classifying, while late fusion fuses the results after classification. The input image is in the space domain. The speech has been converted from the time domain to a log-mel spectrogram, i.e., the space domain. The proposed system contains networks $N1$ and $N2$ and dense, multiply, weighted addition, and softmax layers. $N1$ uses convolution & max-pool layers while $N2$ uses pre-trained networks VGG16 and VGG19 [33]. Both of these networks contain batch-normalization, flattened, and dense layers.

The intuition behind our architecture was to include a mechanism somehow So that each modality affects the other while making predictions. Here the two modalities are combined in two ways:- intermediate fusion and late fusion. First of all, to bring both modalities in the same domain audio signal is converted to a log-mel spectrogram to convert it from the time to space domain. Now, let us consider two networks, N1 and N2. N1 consists of a pre-trained network, than a batch normalization layer, a flattening layer, and a dense layer of 512 neurons. While N2 has the following architecture: First, two convolution layers have 64 filters, then a max-pooling layer, then two more convolution layers of 128 filters, then a max-pooling layer again, comes two more
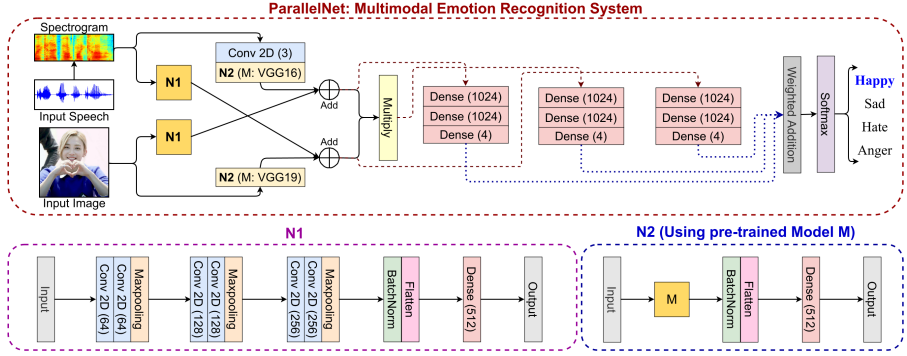
Fig. 2: Architecture of the proposed system (top), $N1$ (bottom left) and $N2$ (bottom right) where M is a pre-trained model.

convolution layers of 256 filters, and then a max-pooling layer. Then it consists of a batch normalization layer, a flattening layer, and a dense layer of 512 neurons.

3.2.1 Intermediate fusion phase

Consider the networks fed with the image input to be $N1_i$ and $N2_i$ while the networks $N1_s$ and $N2_s$ process the speech input. The speech is expressed as a spectrogram of size (128, 128, 1) and passed first to a convolution layer having three convolutional filters of size (1, 1) each and then to $N2_s$ where a pre-trained VGG16 network is used. The image with size (128, 128, 3) is passed to $N2_i$ that uses a pre-trained VGG19 network. As shown in Eq. 1, the output of $N1_s$ is added with the output of $N2_i$ to get F_s. Likewise, the outputs of $N1_i$ and $N2_s$ are added to obtain F_i. Then F_s and F_i are element-wise multiplied to obtain F_{mul}.

$$
\begin{aligned}
F_i &= Add(output(N1_i),\ output(N2_s)) \\
F_s &= Add(output(N1_s),\ output(N1_i)) \\
F_{mul} &= Multiply(F_i,\ F_s)
\end{aligned}
\tag{1}
$$

The choice of using multiplication instead of weighted addition in Eq. 1 to combine F_s and F_i in the low-level fusion has been determined experimentally. Moreover, theoretically, if the speech and image modalities predict the same emotion class, they should support each other. However, let's consider a case where one modality predicts i^{th} emotion very strongly while another predicts another emotion j^{th} weakly. We expect that the ith emotion should be predicted weakly. It would not have been the case in the case of using addition, and the ith emotion would have the upper hand. In comparison, the multiplication of both modalities would dilute the assertive behavior of the ith emotion and give us the expected prediction.
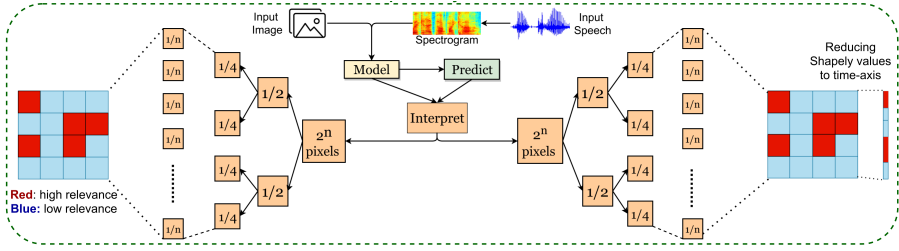
Fig. 3: Proposed interpretability technique's illustration. Here, each part's importance is computed using Divide & Conquer.

3.2.2 Late fusion phase

The intermediate outputs F_i, F_s, and F_{mul} are passed from three dense layers of size 1024, 1024, and 4 to obtain O_{sp} for speech, O_{img} for image, and O_{mul} for multiplied. These outputs are combined using the weighted addition layer as per Eq. 2 in a late fusion manner and passed from a softmax layer to get the final predicted label, $\hat{y}$. The weights w_1, w_2, and w_3 are randomly initialized and passed to a softmax layer to normalize them to non-negative values. Their final values are learned using the Gradient Descent algorithm. It combines the information from speech & image modalities and the output of intermediate fusion in a hybrid manner.

$$O = w_1 \times O_{sp} + w_2 \times O_{img} + w_3 \times O_{mul}$$
$$\hat{y} = Softmax(O)$$
(2)

3.3 Proposed interpretability technique

While making predictions, a deep learning-based classifier is expected to consider the input features that a human would consider. However, it is challenging to look into it and understand what input features it is considering [29]. To work on this challenge, we have developed an interpretability technique based on 'shapely values' [17] that denotes each input feature's importance. Theoretically, shapely values' computation takes exponential time. The computation has been approximated using the divide and conquer approach as shown in Eq. 3. For a model with two features f_1 and f_2, shapely value $\mathscr{S}_{\{f_1\}}$ for feature f_1 denoting its importance is computed as follows.

$$\mathscr{S}_{\{f_1\}} = (1/2) \times MC_{f_1,\{f_1\}} + (1/2) \times MC_{f_1,\{f_1,f_2\}}$$
(3)

Here, $MC_{f_1,\{f_1\}}$ is feature f_1's marginal contribution to the model containing only f_1 and given by Eq. 4 where $score_{\{f_1\}}$ denotes the prediction for the ground-truth label using the model with feature f_1.

$$MC_{f_1,\{f_1\}} = score_{\{f_1\}} - score_{\{\phi\}} \tag{4}$$

The respective speech and image inputs are segregated and fed into the model while keeping the other as zero to compute the individual contribution of each modality. As depicted in Fig. 3, each modality's input is divided into two parts for a specified number of times, and the importance of each part towards the model's prediction is computed as per Eq. 2. Moreover, the calculation of the importance score follows the basic requirement of shapely values given by Eq. 5.

$$\mathscr{S}_{\{f_1\}} + \mathscr{S}_{\{f_2\}} = \mathscr{S}_{\{f_1,\, f_2\}} - \mathscr{S}_{\{null\}} \tag{5}$$

The important image features for the predictions can be directly observed through the shapely values. In contrast, the important speech features are analyzed after transforming them to wave, i.e., time-domain representation. We first applied the shapely values directly and converted the spectrogram to speech; however, the speech reconstructed by this method was not meaningful. Then, we used the method of averaging the shapely values along the frequency axis and reducing them to the time axis to find the features' importance at a given time. The speech segments below a threshold shapely values of 30 percentile have been reduced to zero. The leftover segment with high importance is converted to text using speech-to-text model [3] and interpreted to understand how the model classifies each instance. The proposed interpretability technique has been summarised in Algorithm 1.

4 Experiments

4.1 Experimental setup

The proposed system's network has been trained using Nvidia Quadro P5000 Graphics Card, whereas 64 bit Core(TM) i7-8700 Ubuntu system with 3.70 GHz 16GB RAM has been used for model evaluation.

4.2 Training strategy

The model has been trained using a batch size of 64, a train-test split of 70-30, 5-fold cross-validation, *Adam* optimizer, *ReLU* activation function with a learning rate of 8×10^{-6}. The baselines and proposed models converged regarding validation loss in 18-23 epochs. The models have been trained for 30 epochs as a safe upper bound. A weighted combination of categorical cross entropy with weights 1 and 0.5 and categorical focal loss [16] has been used as the loss function. *EarlyStopping* and *ReduceLROnPlateau* have been incorporated with patience values 5 and 2. Accuracy, macro f1 [23], and *CohenKappa* [38] have been analyzed for evaluation.

Algorithm 1 Proposed interpretability technique

Define $model$: Multimodal deep neural network
Define $data\_img$: Image pixels
Define $data\_speech$: Speech spectrogram pixels
Define $wd,\ ht$: Width & height
Define $times$: Number of division to divide image & speech spectogram in
Define $SHAP\_value\_img$: Image's shaply value
Define $SHAP\_value\_speech$: Speech's shaply value

Procedure DnCShap_MM($model$, $data\_img$, $data\_speech$, wd, ht, $times$)

▷ Initialize data with all zero entries, here; np: numpy
$data\_1$ = np.zeros([$wd,\ ht$, 3])
$data\_2$ = np.zeros([$wd,\ ht$, 1])
$data\_1$ = data_1.reshape(1, $wd,\ ht$, 3)
$data\_2$ = data_2.reshape(1, $wd,\ ht$, 1);

▷ Make original data ready to be fed into $model$
$data\_f\_img$ = $data\_img$.reshape(1, $wd,\ ht$, 3)
$data\_f\_speech$ = $data\_speech$.reshape(1, $wd,\ ht$, 1);

▷ Find the predicted label
$pred$ = $model$.predict ($data\_f\_img$, $data\_f\_speech$)
arg_max = np.argmax($pred$);

▷ Find predicted probability with original data
$pred\_f$ = $pred$[0][arg_max] ;

▷ Find predicted probability with blank data
$pred\_b$ = model.predict($data\_1$, $data\_2$)[0][arg_max];

▷ Find predicted probability with only image modality
$pred\_1$ = model.predict($data\_f\_img$, $data\_2$)[0] [arg_max]

▷ Find predicted probability with only speech modality
$pred\_2$ = model.predict($data\_1$, $data\_f\_speech$)[0] [arg_max];

▷ Compute the importance of image & speech
$score\_1$ = (($pred\_1$ - $pred_b$) + ($pred_f$ - $pred_2$))/2
$score\_2$ = (($pred\_2$ - $pred_b$) + ($pred_f$ - $pred_1$))/2

▷ Placeholders for speech & image shap values
$SHAP\_value\_img$ = np.zeros([$wd,\ ht$])
$SHAP\_value\_speech$ = np.zeros([$wd,\ ht$])
$times$ = $times$ - 1

▷ Compute $SHAP\_value\_img$ and $SHAP\_value\_speech$
Compute $SHAP\_value\_img$ using Eq. 3, 4 and 5
Compute $SHAP\_value\_speech$ using Eq. 3, 4 and 5

4.3 Ablation studies and models

The following studies analyze the effect of using multimodal information and various network configurations.

4.3.1 Effect of multiple modalities

We first worked on SER and IER alone, using only speech samples and images from the IIT-R SIER dataset. Then we combined the information from speech & image modalities and performed multimodal emotion recognition. The IER-only experiments demonstrated high training but low validation accuracy. The convergence of accuracy and f1 score was not in line, and *CohenKappa* metric's value was low, denoting overfitting for a particular class. The accuracy and f1 score converged in line for SER-only experiments, though the accuracy was less.

4.3.2 Effect of various network configurations

As depicted in Fig. 2, ParallelNet consists of a family of networks where $N1$ and $N2$ can be varied in different situations. We first keep $N2$ fixed as EfficientNet [36] and evaluate three configurations for $N1$ – Configuration 1 uses two criss-crosses before and after $N2$. A *criss-cross* is a position combining two different modalities' networks. Configuration 2 & 3 implement single criss-cross before and after $N2$. Three baseline models have been implemented in line with these configurations. Configuration 3 was chosen for final implementation as it shows in-line convergence & improved performance.

Further, keeping Configuration 3 fixed for $N1$'s configuration, following choices have been evaluated for $N2$ – VGG [33] (VGG-16, VGG-19), ResNet [7] (ResNet-34, ResNet-50, ResNet-101), InceptionNet [35] (Inception 3a, Inception 4a), MobileNet [9] and DenseNet [11]. The best performance has been observed with VGG16 as $N2_s$ and VGG19 as $N1_i$, which have finally been implemented by the 'ParallelNet.' The baseline & proposed models determined through the aforementioned studies are listed below, and their performance in terms of validation accuracies have been summarized in Table 2.

- **Baseline 1** – $N1$: Two criss-cross, $N2$: EfficientNet. It divides $N1$ into two parts and uses two criss-crosses before and after $N2$. A *criss-cross* is a position combining two different modalities' networks.

- **Baseline 2** – $N1$: Criss-cross before $N2$; $N2$: EfficientNet.

- **Baseline 3** – $N1$: Criss-cross after $N2$; $N2$: EfficientNet.

- **Proposed** – $N1$: Criss-cross after $N2$; $N2$: VGG.

Table 2: Ablation studies' summary.

Model	Accuracy
SER Only	60.17%
IER Only	66.93%
Baseline 1	63.93%
Baseline 2	61.81%
Baseline 3	67.70%
Proposed ('ParallelNet')	**83.29%**

5 Results and discussion

The emotion classification results have been discussed in this Section, along with their interpretation and a comparison of sentiment classification results with existing methods.

5.1 Quantitative results

The 'ParallelNet' has achieved emotion recognition accuracy of **83.29%**. Its class-wise accuracies are shown in Fig. 4.
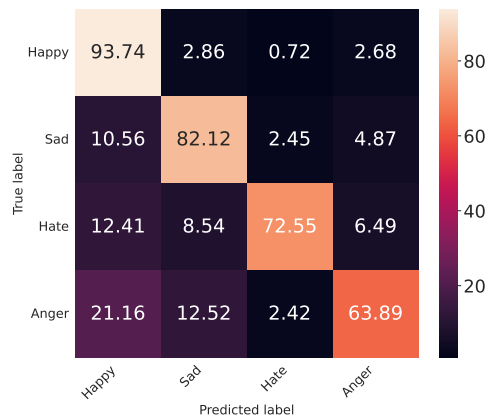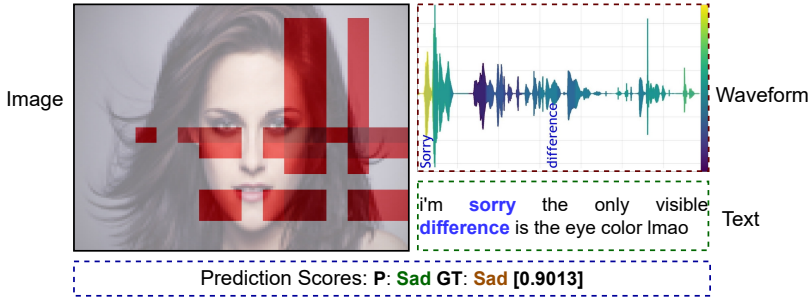

Fig. 4: Confusion matrix showing class-wise accuracies.
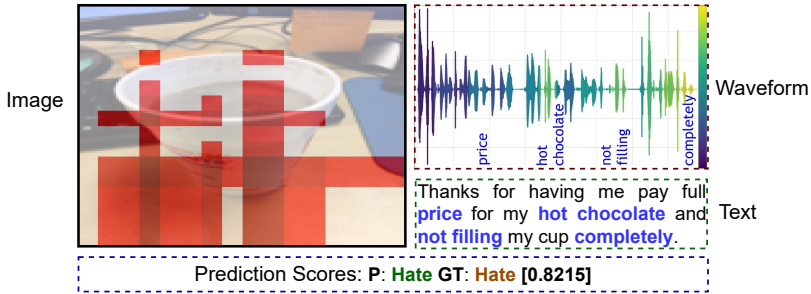
5.2 Qualitative results

Fig. 5 shows sample emotion classification & interpretation results. The important speech and image features contributing to emotion classification are obtained, and corresponding words are highlighted. In the waveform, yellow and blue correspond to the most and least important features.
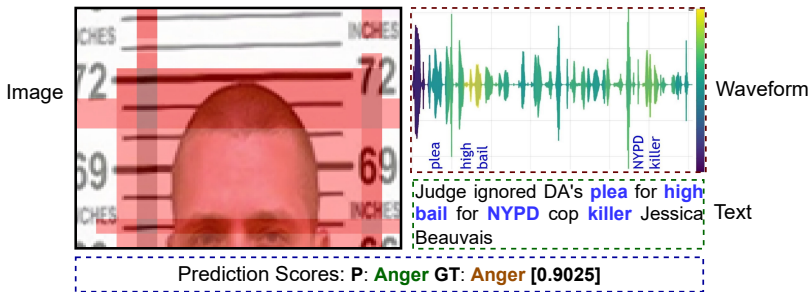
Image   Waveform

birthday   best ever   love u lots

Happy **birthday** to my boo, have the **best** day **ever** and **love u lots** n lots!!!♥♥♥   Text

Prediction Scores: **P**: **Happy** **GT**: **Happy** **[0.9776]**

(a) Sample Result 1

Image   Waveform

Sorry   difference

i'm **sorry** the only visible **difference** is the eye color lmao   Text

Prediction Scores: **P**: **Sad** **GT**: **Sad** **[0.9013]**

(b) Sample Result 2

Image   Waveform

price   hot chocolate   not filling   completely

Thanks for having me pay full **price** for my **hot chocolate** and **not filling** my cup **completely**.   Text

Prediction Scores: **P**: **Hate** **GT**: **Hate** **[0.8215]**

(c) Sample Result 3

Image   Waveform

plea   high bail   NYPD killer

Judge ignored DA's **plea** for **high bail** for **NYPD** cop **killer** Jessica Beauvais   Text

Prediction Scores: **P**: **Anger** **GT**: **Anger** **[0.9025]**

(d) Sample Result 4

Fig. 5: Sample results; here, 'P', 'GT' and 'Score' denote the predicted label, ground-truth label and softmax score.

5.3 Results comparison

Comparison with existing Sentiment Analysis methods: The emotion recognition results have been reported in Section 5.1. The IIT-R SIER dataset has been constructed from the B-T4SA dataset in this paper; hence, there are no existing emotion recognition results for it. However, sentiment classification (into neutral, negative, and positive classes) results on the B-T4SA dataset are available in the literature, which has been compared with the proposed method's sentiment classification results in Table 3.

Table 3: Comparing existing sentiment analysis methods

Approach	Author	Accuracy
Cross-Modal Learning	Vadicamo et al. [37]	51.30%
Multimodal Sentiment Analysis	Gaspar et al. [5]	60.42%
Hybrid Fusion	Kumar et al. [15]	86.70%
ParallelNet (Proposed)		**89.68**%

Comparison with human evaluation: On considering the multimodal context from image and speech modalities, the human evaluation (See Table 1) and automatic evaluation (using ParallelNet. See Table 2) resulted in emotion classification accuracies of 80.46% and 89.68% respectively. In both cases, the emotion classification performance improved on considering the multimodal context compared to considering only image or speech modality. It establishes the importance of considering complementary information from multiple modalities for emotion recognition.

5.4 Discussion

The proposed system classifies a given multimodal input containing speech & the corresponding image into 'anger,' 'happy,' 'hate,' and 'sad' classes. The proposed interpretability technique identifies the important speech & image features contributing to emotion recognition. An alternate procedure to construct the IITR-SIER dataset was to retain only those samples from the BT4SA dataset for which SER & IER models predicted the same label and discard the rest of the samples. However, it would have caused a bias towards the models used in the first place for creating these labels. The SER & IER models have been retrained on the IITR-SIER dataset instead of using the pre-trained weights of the models used to construct the IITR-SIER dataset. However, suppose somebody uses the pre-trained models of either one of the two modalities (trained on IEMOCAP and Flickr & Instagram datasets, respectively) used during dataset construction. In that case, they will get a 100% accuracy. The closest to them, any other evaluated machine learning model is, the more favorable its evaluation would be. That's why the proposed procedure of considering the prediction probabilities for all emotion classes is more effective in capturing the overall emotional context represented by both modalities in combination. It leads to generating more accurate ground-truth labels.

The ParallelNet's architecture has been determined through extensive ablation studies. It consists of a family of networks where $N1$ and $N2$ can be varied in different situations. We've first determined the optimal configuration for $N1$ to combine speech & image modalities' information. Further, VGG, ResNet, InceptionNet, MobileNet, and DenseNet have been evaluated for $N2$. The best performance has been observed with VGG. The ResNet depicted very slow learning for a lower learning rate, while the learning fluctuated significantly for a higher learning rate. The model converged faster for the Inception Net and Efficient Net; however, the accuracy is lower. MobileNet and DenseNet have also resulted in low performance.

Apart from the experimental validation in Table 2, Fig. 5 qualitatively re-affirms the importance of combining complementary information from multiple modalities for more accurate emotion recognition. In the first & second cases, the image and speech features (shown by yellow parts of the waveform and denoted by corresponding words in blue) contribute to predicting the emotion class 'sad.' In the third & fourth cases, the image features have not been precisely captured, and the images seem neutral. However, the corresponding speech features *not filling* and *killer* contribute towards hatred and anger intent, which leads to recognizing the 'hate' and 'anger' classes.

6 Conclusions and future work

The importance of utilizing information from multiple modalities has been established for emotion recognition. The proposed system, *ParallelNet*, has resulted in better performance than SER alone, IER alone, and baseline models. The proposed interpretability technique identifies the important image & speech features contributing to emotion recognition.

Future research plans include working on emotion recognition in other modalities such as text, videos, and emotion signal data. It is also planned to explore the interpretability of emotion recognition in the aforementioned modalities.

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.

Declarations

Conflicts of interest: Authors have no conflict of interest.

Code availability: available at https://github.com/MIntelligence-Group/SpeechImg_EmoRec.

Availability of data and material: available at https://github.com/MIntelligence-Group/SpeechImg_EmoRec.

Authors' contributions: *Puneet Kumar*: Methodology, Implementation, Experiments, Result Analysis, Writing - original draft & editing. *Sarthak Malik*: Data Curation, Implementation, Conceptualization, Validation, Writing - review. *Balasubramanian Raman*: Conceptualization, Writing - review, Supervision, Project administration.

Ethics approval: 'Not applicable'.

Consent to participate: 'Not applicable'.

Consent for publication: 'Not applicable'.

This article does not contain any studies with human participants or animals performed by any of the authors

References

[1] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 41(2):423–443, 2018.

[2] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. IEMOCAP: Interactive Emotional dyadic MOtion CAPture data. *Language Resources and Evaluation*, 42(4), 2008.

[3] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964, 2016.

[4] Dongyang Dai, Zhiyong Wu, Runnan Li, Xixin Wu, Jia Jia, and Helen Meng. Learning Discriminative Features from Spectrograms using Center Loss for SER. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7405–7409, 2019.

[5] António Gaspar and Luís A Alexandre. A Multimodal Approach to Image Sentiment Analysis. In *Springer International Conference on Intelligent Data Engineering and Automated Learning (IDEAL)*, pages 302–309, 2019.

[6] Chen Guanghui and Zeng Xiaoping. Multimodal Emotion Recognition by Fusing Correlation Features of Speech-Visual. *IEEE Signal Processing Letters*, 28:533–537, 2021.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

[8] M Shamim Hossain and Ghulam Muhammad. Emotion Recognition using Deep Learning Approach from Audio Visual Emotional Big Data. *Information Fusion*, 49:69–78, 2019.

[9] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv preprint arXiv:1704.04861*, 2017. Accessed on 06.01.2023.

[10] Anthony Hu and Seth Flaxman. Multimodal Sentiment Analysis to Explore the Structure of Emotions. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pages 350–358, 2018.

[11] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely Connected Convolutional Networks. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017.

[12] Hye-Rin Kim, Yeong-Seok Kim, Seon Joo Kim, and In-Kwon Lee. Building Emotional Machines: Recognizing Image Emotions through Deep Neural Networks. *IEEE Transactions on Multimedia (T-MM)*, 20(11):2980–2992, 2018.

[13] Puneet Kumar, Sidharth Jain, Balasubramanian Raman, Partha Pratim Roy, and Masakazu Iwamura. End-to-End Triplet Loss based Emotion Embedding System for Speech Emotion Recognition. In *IEEE International Conference on Pattern Recognition (ICPR)*, pages 8766–8773, 2021.

[14] Puneet Kumar, Vishesh Kaushik, and Balasubramanian Raman. Towards the Explainability of Multimodal Speech Emotion Recognition. In *INTERSPEECH*, pages 1748–1752, 2021.

[15] Puneet Kumar, Vedanti Khokher, Yukti Gupta, and Balasubramanian Raman. Hybrid Fusion based Approach for Multimodal Emotion Recognition with Insufficient Labeled Data. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 314–318. IEEE, 2021.

[16] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[17] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *The 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pages 4768–4777, 2017.

[18] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. In *Conference on Artificial Intelligence (AAAI)*, volume 33, pages 6818–6825, 2019.

[19] Mariana Rodrigues Makiuchi, Kuniaki Uto, and Koichi Shinoda. Multimodal Emotion Recognition with High-level Speech and Text Features. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021.

[20] Sarthak Malik, Puneet Kumar, and Balasubramanian Raman. Towards Interpretable Facial Emotion Recognition. In *The 12th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, pages 1–9, 2021.

[21] Deep Mind. Wavenet: A Generative Model for Raw Audio. deepmind.com/blog/article/wavenet-generative-model-raw-audio, 2016. Accessed on 20.02.2022.

[22] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A Generative Model for Raw Audio. *arXiv preprint arXiv:1609.03499*, 2016.

[23] Juri Opitz and Sebastian Burst. Macro f1 and Macro f1. *arXiv preprint arXiv:1911.03347*, 2019.

[24] Mathieu Pagé Fortin and Brahim Chaib-draa. Multimodal Multitask Emotion Recognition Using Images, Texts and Tags. In *ACM Workshop on Cross-modal Learning and Application*, pages 3–10, 2019.

[25] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. DeepVoice 3: Scaling Text-to-Speech with Convolutional Sequence Learning. In *The 6th Int. Conference on Learning Representations (ICLR)*, 2018.

[26] Robert Plutchik. The Nature of Emotions. *Journal Storage Digital Library's American scientist Journal*, 89(4):344–350, 2001.

[27] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion. *Elsevier Information Fusion Journal*, 37:98–125, 2017.

[28] Tianrong Rao, Xiaoxu Li, and Min Xu. Learning Multi-level Deep Representations for Image Emotion Classification. *Neural Processing Letters*, pp. 1–19, 2019.

[29] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why Should I Trust You? Explaining Predictions of Any Classifier. In *International Conference on Knowledge Discovery & Data mining (KDD)*, pages 1135–1144, 2016.

[30] Justin Salamon and Juan Pablo Bello. Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. *IEEE Signal Processing Letters*, 24(3):279–283, 2017.

[31] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *The IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[32] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning Important Features Through Propagating Activation Differences. In *International Conference on Machine Learning (ICML)*, pages 3145–3153, 2017.

[33] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556*, 2014. Accessed on 06.01.2023.

[34] Shamane Siriwardhana, Andrew Reis, and Rivindu Weerasekera. Jointly Fine Tuning 'BERT-Like' Self Supervised Models to Improve Multimodal Speech Emotion Recognition. *INTERSPEECH*, pages 3755–3759, 2020.

[35] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[36] Mingxing Tan and Quoc Le. EfficientNet: Rethinking Model Scaling for CNN. In *International Conference on Machine Learning (ICML)*, pages 6105–6114, 2019.

[37] Lucia Vadicamo, Fabio Carrara, Andrea Cimino, Stefano Cresci, Felice Dell'Orletta, Fabrizio Falchi, and Maurizio Tesconi. Cross-Media Learning for Image Sentiment Analysis in the Wild. In *IEEE International Conference on Computer Vision Workshops (ICCV-W)*, pages 308–317, 2017.

[38] Susana M Vieira, Uzay Kaymak, and João MC Sousa. Cohen's Kappa Coefficient as a Performance Measure for Feature Selection. In *International Conference on Fuzzy Systems*, pages 1–8. IEEE, 2010.

[39] Mingke Xu, Fan Zhang, and Samee U Khan. Improve Accuracy of Speech Emotion Recognition with Attention Head Fusion. In *IEEE Annual Computing and Communication Workshop and Conference (CCWC)*, pages 1058–1064, 2020.

[40] Promod Yenigalla, Abhay Kumar, Suraj Tripathi, Chirag Singh, Sibsambhu Kar, and Jithendra Vepa. Speech Emotion Recognition Using Spectrogram &

Phoneme Embedding. In *INTERSPEECH*, pages 3688–3692, 2018.

[41] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Building a Large Scale Dataset for Image Emotion Recognition: the Fine Print and the Benchmark. In *The 30th AAAI Conference on Artificial Intelligence (AAAI)*, pages 308–314, 2016.

[42] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A Survey of Affect Recognition: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 31(1):39–58, 2009.