This paper proposes a multimodal emotion recognition system, VIsual Spoken Textual Additive Net (VISTA Net), to classify the emotions reflected by a multimodal input containing image, speech, and text into discrete classes. A new interpretability technique, K-Average Additive exPlanation (KAAP), has also been developed to identify the important visual, spoken, and textual features leading to predicting a particular emotion class. The VISTA Net fuses the information from image, speech & text modalities using a hybrid of early and late fusion. It automatically adjusts the weights of their intermediate outputs while computing the weighted average without human intervention. The KAAP technique computes the contribution of each modality and corresponding features toward predicting a particular emotion class. To mitigate the insufficiency of multimodal emotion datasets labeled with discrete emotion classes, we have constructed a large-scale IIT-R MMEmoRec dataset consisting of real-life images, corresponding speech & text, and emotion labels (‘angry,’ ‘happy,’ ‘hate,’ and ‘sad.’). The VISTA Net has resulted in 95.99% emotion recognition accuracy on considering image, speech, and text modalities, which is better than the performance on considering the inputs of any one or two modalities.

CCS Concepts: • Information systems → Sentiment analysis; Multimedia and multimodal retrieval; • General and reference → Cross-computing tools and techniques; • Computing methodologies → Supervised learning by classification.

Additional Key Words and Phrases: Affective Computing, Emotional and Social Signals, Multimodal Fusion, Media Interpretation, Emotion and Sentiment Analysis.

ACM Reference Format:
Puneet Kumar, Sarthak Malik, and Balasubramanian Raman. 2022. Hybrid Fusion Based Interpretable Multimodal Emotion Recognition with Insufficient Labelled Data. ACM Trans. Multimedia Comput. Commun. Appl. 1, 1, Article 1 (July 2022), 23 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The multimedia data has overgrown in the last few years, leading multimodal emotion analysis to emerging as an important research trend [2]. 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 [49]. Humans portray different emotions through various modalities such as images, speech, and text [8]. Utilizing the multimodal information from them could increase the performance of emotion recognition [75].

*denotes Equal Contribution

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2022 Association for Computing Machinery.
1551-6857/2022/7-ART1 $15.00
https://doi.org/XXXXXXX.XXXXXXX
Researchers have performed emotion recognition by analyzing visual, spoken, and textual information separately [38, 53, 69]. The multimodal emotion recognition using two of these modalities has also been explored; however, emotion recognition using all three modalities, i.e., visual, spoken, and textual, has not been fully explored [75]. 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 visual, spoken, and text modalities and explaining each modality’s importance and particular features contributing to emotion recognition.

Multimodal emotion recognition also faces the unavailability of sufficient labeled data 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 [5]. A few multimodal datasets contain generic images; however, they consist of positive, negative, and neutral sentiment labels instead of multi-class emotion labels [16, 67]. In this chapter, we construct a new dataset, the IIT-R MMEmoRec dataset, that contains generic images, corresponding speech utterances, text transcripts, and discrete class labels, i.e., ‘happy,’ ‘sad,’ ‘hate,’ and ‘anger.’

We start with bimodal emotion recognition using the pre-trained SER, TER, and IER models build in the previous chapters. Intermediate fusion is first implemented on two modalities’ information, then late fusion is applied to the intermediate fusion’s output and respective emotion recognition modules’ output. For instance, to perform (speech + text) emotion recognition, the intermediate fusion of speech and text information is performed, whose output is later fused with the outputs of SER and TER models. The modality weights for the fusion are computed using the grid search. Further, we propose an interpretable multimodal emotion recognition system, VISTA Net, that fuses image, speech, and text features using hybrid fusion. A novel interpretability technique, KAAP, has also been developed to identify the important visual, spoken, and textual features that predict particular emotion classes. The VISTA Net uses KAAP and automatically adjusts the weights of their intermediate outputs while computing the weighted average without human intervention.

Using the modality weights computed through grid-search, 82.63% accuracy has been observed for (speech + text) emotion recognition whereas (image + text) emotion recognition and (speech + image) emotion recognition resulted in 90.20% and 84.17% accuracies on IIT-R MMEmoRec dataset. The VISTA Net has resulted in 95.99% emotion recognition accuracy when considering the image, speech, and text modalities, which is better than the performance when considering the input of any one or two modalities. The enhanced performance by the VISTA Net advocates the importance of utilizing complementary information from multiple modalities for emotion recognition. The importance of all the modalities has been reported, and their features contributing to emotion recognition are highlighted during the analysis. This paper makes the following contributions.

1. The bimodal emotion recognition has been performed for each combination of speech, text, and image modalities using a hybrid of intermediate and late fusion. The SER, TER, and IER models proposed in the previous chapters have been utilized, and modality weights for fusion have been computed using grid-search.

2. A hybrid-fusion-based novel interpretable multimodal emotion recognition system, VISTA Net, has been proposed to classify an input containing an image, corresponding speech, and text into discrete emotion classes.

3. A novel interpretability technique, KAAP, has been developed to identify each modality’s importance and important image, speech, and text features contributing the most to recognizing emotions.

4. A large-scale dataset, ‘IIT-R MMEmoRec dataset’ containing images, speech utterances, text transcripts, and emotion labels has been constructed.
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 construction procedure. Section 4 and 5 discuss the experiments and results and the paper is concluded in Section 6.

2 RELATED WORKS

2.1 Unimodal emotion recognition

2.1.1 Speech emotion recognition. The traditional feature-based SER systems extract audio features such as cepstrum coefficient, voice tone, prosody, and pitch and use them for SER [13]. For instance, Rong et al. [55] worked on extracting the most important audio features from the speech samples, whereas Lee et al. [30] used the extracted audio features to identify negative and positive emotions in speech samples. The feature-based SER systems depend on the polarity of the emotional features. The features of high-key classes (happiness and anger) are similar in properties among themselves, and they are very different from the low-key classes (sad and despair) [64]. In the context of machine learning-based SER, Support Vector Machine (SVM) based classifiers and Hidden Markov model (HMM) based statistical techniques have also been explored [23, 34]. However, they require manual crafting of acoustic features, and HMM-based models cannot always reliably estimate the parameters of global speech features [13]. Hence, it is challenging to develop an end-to-end SER system using them. The deep learning-based approaches using spectrogram features and attention mechanisms have shown state-of-the-art SER results [10]. In this context, Xu et al. [69] generated multiple attention maps, fused them, and used them for SER. They observed an increased performance as compared to non-fusion-based approaches. In another work, Mao et al. [41] used a Convolutional Neural Network (CNN) for spectrogram processing. On the other hand, Majumder et al. [38] performed Recurrent Neural Network (RNN) based SER. They determined speech embeddings and used them for speaker identification. In another work using attention maps, Seyedmahdad et al. [43] implemented local attention to learn the emotion features automatically.

2.1.2 Text emotion recognition. Deep learning-based approaches have shown state-of-the-art TER results. With the evolution of deep learning, it has become possible to convert text into vectors and use Deep Neural Networks (DNNs) to process them. Emotion recognition in conversation could be useful to mine opinions from conversational data on platforms such as YouTube, Facebook, Reddit, Twitter, and others [51]. More examples of Deep Learning-based emotion analysis include personality detection from text using document modeling [37] and text-based emotion recognition using YouTube comments [19]. In another work, Huang et al. [22] used a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model for emotion recognition in text dialogs. Shrivastava et al. [60] applied sequence-based CNNs, whereas Bambaataa et al. [3] utilized semantic and emotional information to train a deep TER model. Deep models focus on learning parameters such as features, sequence-related information, and contextual information from the text input.

2.1.3 Image emotion recognition. One of the most informative ways for machine perception of emotions is through facial expressions in images and videos. Identifying human emotions from facial expressions is a relatively more saturated research area than emotion recognition in other modalities. Various techniques such as face localization, face registration, micro-expression analysis, tracking the landmark points, shape feature analysis, eye gaze prediction, face segmentation, and detection have been developed for facial emotion recognition [9, 31]. Image Emotion Recognition (IER) research is also an active domain. For instance, Kim et al. [? ] 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. [53] prepared hierarchical notations for emotion recognition in the
visual domain. Traditional, feature-based IER analysis with low-level (shape, color, and edge) and mid-level (composition and optical balance) image features [20, 24, 36] used the semantic content of the images for emotion analysis. However, all the low and mid-level features are difficult to accommodate by handcrafted feature extraction techniques used for machine traditional and machine learning-based methods. On the other hand, deep learning-based IER approaches are capable of extracting high-level visual features, but they struggle to extract low and mid-level features. Moreover, they require well-labeled large-scale datasets for training [53].

2.2 Multimodal emotion recognition
Different emotion representation methods are used in various modalities [18, 71]. For example, emotion representation in face and gestures utilizes features tracking, sensitivity analysis, and heat maps. In real-life scenarios, emotions are portrayed through various modalities such as vision, speech, and text. Analysis in a single modality may not be able to recognize the emotional context completely [75]. That fact has paved the way to draw researchers’ attention towards multimodal emotion analysis [21]. Moreover, various modalities have different statistical properties associated with them. To correctly recognize complex human emotions portrayed through them, it is very important to consider their inter-relationships [49]. As discussed earlier in this chapter, various attempts have been made for emotion analysis using visual, spoken, and textual modalities individually. However, the emotion analysis in a multimodal manner considering the inter-relationships of these modalities is still an unexplored space [46]. The existing works in this direction are discussed as follows.

2.2.1 (Speech + text) emotion recognition. In the context of recognizing emotions from speech utterances and corresponding text transcripts, Chuang et al. [7] worked on analyzing the overlapping emotional information in speech and text. A simple fusion of spoken and textual information has also been used for emotion recognition in some more works. For instance, Makuuchi et al. [39] performed separate acoustic and textual analyses and determined the emotional context based on their collective result. In another work, Yoon et al. [72] extracted the audio and text information using dual RNNs and then combined it to perform emotion recognition. On the other hand, some research attempts used textual information to improve the SER performance. For example, Tripathi et al. [65] performed emotion recognition on the Interactive EMOTIONal dyadic motion CAPturer (IEMOCAP) dataset [5] using data from speech and text modalities and complementing it with hand movements and facial impressions data. In another work, Siriwardhana et al. [63] fine-tuned transformers-based models to improve the performance of multimodal speech emotion recognition.

2.2.2 (Text + image) emotion recognition. Several attempts have been made to recognize the emotional content portrayed in visual and textual modalities. In this direction, Kahou et al. [25] developed a framework, EmoNets for emotion recognition in video and text. Fortin et al. [46] implemented a multi-task architecture-based emotion recognition approach to perform predictions with one or two missing modalities by using a classifier for each combination of image, text, and tags. In another work, Xu et al. [70] modeled the interplay of visual and textual content for sentiment recognition using a co-memory based-network.

2.2.3 (Speech + image) emotion recognition. Multimodal emotion analysis from audio-visual data has also started getting researchers’ attention lately [21]. For instance, Aytar et al. [1] proposed SoundNet that extracts the emotional information by self-supervised learning of sound representations. In another work, Guanghui et al. [17] implemented the feature correlation analysis algorithm for multimodal emotion recognition. They extracted speech and visual features using two-dimensional and three-dimensional CNNs, fused the features, and used SVM for emotion classification on the fused features.
2.2.4 (Speech + text + image) emotion recognition. There have been several attempts regarding emotion recognition in more than two modalities simultaneously. For example, Poria [50] used information fusion techniques to combine the context from audio, vision, and textual modalities for the sentiment analysis. They found the fusion of the textual and spoken description of the emotional information as an aid to the emotion analysis for visual modality. In another work, Tzirakis et al. [66] extracted the features from speech, image, and text modalities, analyzed the correlation among them, and trained an end-to-end emotion recognition system using them combinedly.

2.3 Explainable and interpretable emotion analysis

Explainability refers to the ability to describe an algorithm’s mechanism that led to a particular output. In contrast, Interpretability understands the context of a model’s output, analyzes its functional design, and relates the design to the output [4, 28]. The deep learning-based techniques act like a black box. The challenges involved with explaining and interpreting their internal working have given rise to a new research area known as explainable AI [35]. Among the recent research carried out in this direction, Riberio et al. [54] pointed out the value of interpreting the internal working of deep learning-based classifiers. They also designed a framework that computes the importance of each input towards a particular output and interprets a classifier’s predictions. In another research, a method to determine the input’s part leading to a particular output has been developed by Fazi et al. [14]. Research has also been done for tracing every neuron’s contribution and understanding the output part-by-part [59]. The existing Interpretability techniques can be divided into the following categories.

2.3.1 Attribution Interpretability techniques. In these methods, the attribution values denoting the relevance of inputs concerning outputs are determined. A popular attribution value is ‘Shapley Values’ [58]. The attribution techniques are frequently used for local interpretability, which explains the impact of one instance instead of the overall model. The Shapley values have been used by Lundberg et al. [35] who implemented an interpretability framework, SHAP (Shapley Additive exPlanations), which determines each feature’s contribution by analyzing its Shapley values [40]. The computation of Shapley values is very expensive because $2^n$ models are required to be trained for a model with $n$ features [6, 35]. Different approximations have been used to speed us Shapley values’ computation. For instance, Shapley values sampling [6] and KernelSHAP [35]. The attribution techniques are further classified into perturbation and back-propagation-based approaches, which are explained further.

2.3.2 Perturbation Interpretability techniques. These techniques make a small change in the input and observe its impact. The insights thus obtained are used to interpret the model’s working [15]. The most frequently used perturbation technique is Local Interpretable Model-agnostic Explanations (LIME) [54] that perturbs the given instance and synthesizes new data. The new data is weighted as per the closeness of the new instance to the original instance. The output is computed by training the original model on the perturbed data. The trained model’s weights denote the approximate values of each feature’s contribution. The LIME can be used with any machine learning model, though it is computationally expensive as it involves the generation of new data.

2.3.3 Backpropagation Interpretability techniques. The backpropagation-based interpretability techniques calculate the attributions by backpropagating through the network multiple times. A popular backpropagation-based technique is ‘Saliency Map’ [61] that has the label output’s absolute gradient of each input feature as an attribution. Another popular technique is Gradient-weighted Class Activation Map (Grad-CAM), which assigns a score to each feature and computes the activation map using this score [57]. The Grad-CAM goes until the last convolutional layer instead of
backpropagating back to the image. It generates a map that highlights the important features of the input image. If the input image is slightly changed, it generates an entirely different map [27].

As suggested by the above survey, LIME, SHAP, and Grad-CAM are the most frequently used interpretability techniques for machine learning models. The LIME is involved with a very high computational cost, whereas Grad-CAM is incapable of withstanding small changes in the input image. The SHAP technique does not suffer from the aforementioned limitations. Furthermore, the DNN interpretability has been applied to the visual modality but it has not been fully explored for the speech and text modalities and multimodal analysis. It inspired us to develop an interpretability technique for multimodal emotion recognition to explain the importance of each modality and identify the important features of each modality that lead to the prediction of a particular class.

3 PROPOSED WORK

3.1 Data compilation

Table 1 shows some samples from the IIT-R MMEmoRec dataset while the process to construct it has been elaborated as follows. It contains generic (facial, human, non-human objects) images (as opposed to only facial images/videos in other known trimodular emotion datasets, IEMOCAP [5], and MOSEI [74]), speech utterances, text transcripts, emotion label (‘angry,’ ‘happy,’ ‘hate,’ and ‘sad’), the probability of each emotion class given by each modality and probability of final emotion class. The IIT-R MMEmoRec dataset has been constructed on top of the ‘Balanced Twitter for Sentiment Analysis’ (B-T4SA) dataset [67]. The B-T4SA dataset contains images, text, and sentiment (‘positive,’ ‘negative,’ neutral) labels, whereas the IIT-R MMEmoRec dataset, has been compiled to have discrete emotion labels for image, text, and speech modalities. The following steps have been followed to construct the IIT-R MMEmoRec dataset.

Table 1. A few samples from IIT-R MMEmoRec dataset. Here, ‘Img_Prob,’ ‘Sp_Prob,’ ‘Txt_Prob,’ and ‘Final_Prob’ are image, speech, text and final prediction probabilities whereas angry, happy, hate and sad emotion labels are denoted as 0, 1, 2 & 3 respectively.

| Image | Speech/Spectrogram | Text | Img_Prob | Sp_Prob | Txt_Prob | Final_Prob | Label |
|-------|---------------------|-------|----------|---------|----------|------------|-------|
| ![Image](example.jpg) | Are These the black lives thugs you are talking about? | ![Speech](example.wav) | 0: 0.7812 | 0: 0.3740 | 0: 0.9204 | 0: 0.6919 | 0 (Angry) |
| ![Image](example.jpg) | Great start to the senior year! | ![Speech](example.wav) | 0: 0.2584 | 0: 0.0245 | 0: 0.0038 | 0: 0.0956 | 1 (Happy) |
| ![Image](example.jpg) | Unbelievable images of dead mermaid washed up on beach, see shocking photos. | ![Speech](example.wav) | 0: 0.0030 | 0: 0.0833 | 0: 0.9009 | 0: 0.3318 | 2 (Hate) |
| ![Image](example.jpg) | Oh SJWs, your tears say more than your nonsensical drivel ever could. #StaySalty #CryMeARiver | ![Speech](example.wav) | 0: 0.1777 | 0: 0.0414 | 0: 0.9447 | 0: 0.3879 | 3 (Sad) |

- The text from the BT4SA dataset is pre-processed by removing links, special characters, and tags, and then the cleaned text is converted to speech using the pre-trained state-of-the-art text-to-speech (TTS) model, DeepSpeech3 [47]. The rationale for using the TTS model is...
Hybrid Fusion Based Interpretable Multimodal Emotion Recognition with Insufficient Labelled Data

governed by the recent studies that prove TTS models generate high-quality speech signals that can be used as a valid approximation of natural speech signals [42, 44, 47].

- The image, speech, and text components are passed through pre-trained IER, SER, and TER models trained on Flickr & Instagram (FI) [73] dataset, IEMOCAP [5] dataset, and ISEAR dataset [56], respectively, and the prediction probabilities of each emotion class are obtained for each modality.
- The prediction probabilities are then averaged to obtain the ground-truth emotion of each data sample. The averaging is done to ensure that the chosen ground truth is the one that is supported by the majority of modalities. Fig. 1 shows an example of emotion label determination. The probabilities for each emotion class given by each modality are shown. The ‘happy’ class has an average prediction probability of 0.50 compared to 0.233 for ‘angry,’ 0.133 for ‘hate,’ and 0.133 for ‘sad.’ The final emotion label for the sample is determined as ‘happy.’

| Emotion | Samples |
|---------|---------|
| Angry   | 53,317  |
| Happy   | 44,980  |
| Hate    | 3,831   |
| Sad     | 10,327  |

Table 2. Class-wise data distribution.

- The data is segregated according to classes, and the samples having an average prediction probability of less than the threshold confidence value of 0.55 times of maximum probability for the corresponding class are discarded. The threshold confidence is determined in Section 4.3.2.
- The four emotion classes, ‘angry,’ ‘happy,’ ‘hate,’ and ‘sad,’ are common in various datasets of different modalities considered in this work. The samples labeled as ‘excitement’ & ‘disgust’ have been re-labeled as ‘happy’ & ‘hate’ as per Plutchik’s wheel of emotions [48]. The final dataset contains a total of 1,12,455 samples with 53,317 labeled as ‘angry,’ 44,980 as ‘happy’ and 10,327 & 3,831 as ‘sad’ and ‘hate’ respectively as described in Table 2.

3.1.1 Determining threshold confidence value for dataset construction. The original BT4SA dataset contained 4.7M data samples labeled as ‘positive,’ ‘negative,’ and ‘neutral.’ While constructing the IIT-R MMEmoRec dataset with discrete emotion labels, i.e., ‘angry,’ ‘happy,’ ‘hate,’ and ‘sad,’ it was essential to retain only the samples having high confidence in the associated emotion label. After passing the image, speech, and text components of the inputs to respective emotion recognition models as discussed in Section 3.2, we computed a value for each data sample in each class representing at what percentage compared to the class maximum that sample is in its particular class. It gave us the confidence of each data sample in its particular class. To determine the appropriate threshold, we plotted possible threshold values vs. the ratio of the class present (the number of each class sample and the total number of samples) as shown in Fig. 2.

The higher the threshold, the higher the confidence and the better the quality of data. However, a higher threshold value also leads to two issues – i) reduction in the size of the dataset and ii) disruption in the distribution of emotion classes compared to its original distribution. As seen in Fig. 2 the distribution of each class at a threshold approaching 1 is very different as compared to when all samples are taken at the 0 thresholds.
An appropriate threshold value needs to be chosen, leading to a good trade-off between high confidence and appropriate size & distribution of the dataset. Till the threshold value of 0.33, the distribution is almost the same as the original, but this confidence is too low to be acceptable. The next possible value is above 0.5 but below 0.6. Between these two values, the distribution of various classes is almost the same, and the confidence is also above 0.5, which is acceptable. Hence an average value of 0.55 is chosen as the threshold confidence value.

### 3.1.2 Human evaluation.

The MMEmoRec dataset has been evaluated by having 8 people evaluate the data samples. 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 speech against the human speech recorded by the human readers and scored the contextual similarity between them on a scale of 0 to 100. The human evaluators also evaluated whether the data samples’ speech, image, and text components agree with the annotated emotion sample individually and combinedly. The samples have been picked randomly, and the average of the evaluators’ scores has been reported in Table 3 where $S_{ss-hs}$ denotes the percentage of evaluators reporting the synthetic speech (ss) to be similar to human speech (hs). $S_{ss}$ & $S_{hs}$ denotes the percentage of speech components of synthetic and human speech portraying the annotated emotion. Likewise, $S_i$ and $S_t$ denote the agreement of annotated emotion class by image and text components. $S_{ss-i-t}$ and $S_{hs-i-t}$ show the samples showing agreement of the annotated emotion class by all three modalities on considering synthetic and human speech, respectively.

We had two readers read the text of the data samples and called their output human synthesized speech. 60.72% evaluators found the synthetic speech to be contextually similar to the human synthesized speech. 74.49% synthetic speech samples and 78.91% human synthesized speech samples were found to be portraying the annotated emotion labels. As per the further observations, 69.26% images and 78.81% text components of the data samples correspond to the annotated emotion labels. Moreover, the evaluators also reported that 72.99% of the samples considering machine synthesized speech along with the corresponding text & image were in line with the determined emotion label, whereas this is comparable to the value of 76.74% on considering human synthesized speech along with the corresponding text & image.
Table 3. Human evaluation of MMEmoRec dataset.

| Class | \(S_{ss-hs}\) | \(S_{ss}\) | \(S_{hs}\) | \(S_i\) | \(S_{ss-i-t}\) | \(S_{hs-i-t}\) |
|-------|---------------|----------|----------|-------|----------------|----------------|
| Angry | 67.18%        | 82.81%   | 85.94%   | 70.31%| 84.38%         | 75.00%         |
| Happy | 52.78%        | 66.67%   | 69.44%   | 63.89%| 72.22%         | 66.67%         |
| Hate  | 62.50%        | 71.43%   | 72.32%   | 67.86%| 73.21%         | 72.32%         |
| Sad   | 60.42%        | 77.08%   | 78.13%   | 75.00%| 87.50%         | 77.08%         |
| Overall| 60.72%        | 74.49%   | 76.46%   | 69.26%| 78.81%         | 72.99%         |

3.2 VISTA Net

The proposed system, VISTA Net’s architecture, is shown in Fig. 3 which has been decided based on the ablation studies discussed in Section 4.3.1. It fuses image, speech & text features using a hybrid of two-stage intermediate and late fusion, which considers all possible pairs of all three modalities and automatically weights them without human intervention. The intermediate fusion combines the information from various modalities before classifying, i.e., after feature extraction, whereas late fusion combines the information after classification.

![VISTA Net: Visual Speech Textual Additive Net](image)

Fig. 3. Schematic architecture of the proposed multimodal emotion recognition system. Here, \(P_m\) & \(S_m\) denote the pre-trained & simpler networks for \(m^{th}\) modality whereas ‘i,’ ‘s,’ and ‘t’ denote visual, speech and text modalities, respectively.

The three modalities are fed into two types of networks: a pre-trained and a simpler network. The intuition behind this approach is to build a fully automated multimodal emotion classifier by including various modalities’ in all possible combinations and learning their weights while training without any human intervention. The proposed system contains \(P_i\) and \(S_i\) for image, \(P_s\) and \(S_s\) for speech, and \(P_t\) and \(S_t\) for text, denoting pre-trained and simpler networks respectively. The input speech has been converted to a log-mel spectrogram before feeding into the network.

3.2.1 Intermediate fusion phase. The images of dimension \((128, 128, 3)\) are fed into \(P_i\) and \(S_i\) respectively, with \(P_i\) consisting of VGG16 [62] and a 512-dimensional dense layer and \(S_i\) containing 3 convolution layers of 64, 128 and 256 filters of size \((3, 3)\) and a dense layer of 512 dimensions. The spectrogram of size \((128, 128, 1)\) from speech input is passed from a 3 filter convolution layer of
size (1, 1), to make it compatible with VGG16. Further, it is passed from \(P_s\) and \(S_s\), consisting of the same architecture as \(P_i\) and \(S_i\) respectively.

The text input is similarly passed from \(T_i\) containing a BERT [12] and \(T_s\) consisting of an embedding and LSTM layer with 64 units. Both \(T_i\) & \(T_s\) are followed by 512-dimensional dense layers. In the intermediate fusion, all pairs of the pre-trained and simpler networks from different modalities are created by passing them from the WeightedAdd layer that we have defined. It gives us six such combinations passed from 2 dense layers with 1024 neurons, giving the classification based on each pair. The Eq. 1 shows all the possible pairs formed from the combination of pre-trained and simpler networks such that both the networks do not belong to the same modality.

\[
\begin{align*}
O_1 &= WeightedAdd(P_i, S_s) \\
O_2 &= WeightedAdd(P_i, S_t) \\
O_3 &= WeightedAdd(P_s, S_i) \\
O_4 &= WeightedAdd(P_s, S_t) \\
O_5 &= WeightedAdd(P_t, S_i) \\
O_6 &= WeightedAdd(P_t, S_s)
\end{align*}
\]  

(1)

where \(O_1, O_2, O_3, O_4, O_5\) and \(O_6\) are the classification outputs for various pairs of pre-trained and simpler networks. The WeightedAdd layer ensures that during training, the weight of any weighted addition is learned using back-propagation without any human intervention. Each weight in the WeightedAdd layer is randomly initialized and then passed from the softmax layer, giving us positive values used as final weights and learned during training.

3.2.2 Late Fusion Phase. In this phase, the information from various modalities’ all possible pairs is combined in a hybrid manner. The intermediate classification outputs obtained from above Eq. 1 are passed from another WeightedAdd layer, which combines these outputs dynamically, giving us the final output \(O\) as depicted in Eq. 2. The output \(O\) is passed from a dense layer with dimensions equal to the number of emotion classes, i.e., four.

\[
O = WeightedAdd(O_1, O_2, O_3, O_4, O_5, O_6)
\]  

(2)

where \(O\) denotes the final output and \(O_1, O_2, O_3, O_4, O_5\) and \(O_6\) are the intermediate classification outputs.

3.3 KAAP

This Section proposes a novel multimodal interpretability technique, K-Average Additive exPlanation (KAAP), depicted in Fig. 4. It computes the importance of each modality and its features while predicting a particular emotion class. The existing interpretability techniques do not apply to speech and multimodal emotion recognition. Moreover, the most frequently used and accepted interpretability technique for images and text is SHAP [35], which is an approximation of Shapley values [58]. It requires \(O(n^2)\) computational time-complexity whereas KAAP requires a time of \(O(k^2)\) where \(k <= n\) is a given hyper-parameter. Moreover, KAAP applies to multimodal emotion analysis and a single modality or a combination of any two modalities as well.

3.3.1 Calculating K exPlanable (KP) values. For a model with \(k\) features \(\{f_1, f_2, \ldots f_k\}\), K exPlanable (KP) value of feature \(f_i\), \(kp_{f_i}\) denotes its importance. Fig. 5 depicts an example calculation.

Consider four nodes, Node 1 with no feature i.e NULL, Node 2 with a single feature \(f_i\), Node 3 with all the remaining features left in Node 1, i.e., \(\{f_1, f_2, \ldots, f_{(i-1)}, f_{(i+1)}, \ldots, f_k\}\), and Node

ACM Trans. Multimedia Comput. Commun. Appl., Vol. 1, No. 1, Article 1. Publication date: July 2022.
Fig. 4. Schematic representation of the proposed interpretability technique. Where $k_i$, $k_s$, and $k_t$ denote the no. of partitions for image, speech and text modalities; $w_i$, and $w_s$ are the widths for image & speech feature matrices and $L_t$ is text feature vector’s length.

**Algorithm 1:** Predict (Probability prediction)

Define model: DNN model,
Define img: Image data,
Define text: text data,
Define speech: Speech spectogram data,
Define type: The type of data whose perturbed probability is required

procedure Predict (model, img, text, speech, w, type)
    if type == image then
        prob = model.predict_proba(img, 0, 0)
    end
    if type == text then
        prob = model.predict_proba(0, text, 0)
    end
    if type == speech then
        prob = model.predict_proba(0, 0, speech)
    end
    return prob

Fig. 5. Sample model for KP values computation.
**Algorithm 2: KAAP_helper** (Procedure to calculate KAAP value for each data instance)

Define `model`: DNN model,  
Define `img`: Image data,  
Define `text`: text data,  
Define `speech`: Speech spectogram data,  
Define `w`: Size of data  
Define `k`: Parts in which data is divided  
Define `type`: The type of data KAAP values are required

procedure KAAP_helper (`model, img, text, speech, w, k, type`)

\[ p_f = \text{Predict}(\text{img, text, speech, type}) \]

\[ p_b = \text{Predict}(0, 0, 0, \text{type}) \]

\[ \text{kaap} = \text{np.zeros}(w) \]

for \( i \leftarrow k \) do

\[ \text{parts} = \text{make_parts}(w, i) \]

for \( j \leftarrow \text{len(parts)} \) do

\[ \text{part} = \text{parts}[j] \]

if \( \text{type} == \text{image} \) then

\[ \text{data}_1 = \text{np.zeros}(w) \]

\[ \text{data}_1[\text{part}] = \text{img}[\text{part}] \]

\[ \text{data}_2 = \text{img} \]

\[ \text{data}_2[\text{part}] = \text{np.zeros}(\text{part}) \]

end

if \( \text{type} == \text{text} \) then

\[ \text{data}_1 = \text{np.zeros}(w) \]

\[ \text{data}_1[\text{part}] = \text{text}[\text{part}] \]

\[ \text{data}_2 = \text{text} \]

\[ \text{data}_2[\text{part}] = \text{np.zeros}(\text{part}) \]

end

if \( \text{type} == \text{speech} \) then

\[ \text{data}_1 = \text{np.zeros}(w) \]

\[ \text{data}_1[\text{part}] = \text{speech}[\text{part}] \]

\[ \text{data}_2 = \text{speech} \]

\[ \text{data}_2[\text{part}] = \text{np.zeros}(\text{part}) \]

end

\[ p_1 = \text{Predict}(\text{data}_1, \text{data}_1, \text{data}_1, \text{type}) \]

\[ p_2 = \text{Predict}(\text{data}_2, \text{data}_2, \text{data}_2, \text{type}) \]

\[ kp\_values = \frac{1}{2} \ast (p_1 - p_b) + \frac{1}{2} \ast (p_f - p_2) \]

if \( \text{type} == \text{image} \) or \( \text{type} == \text{speech} \) then

\[ \text{kaap} += \left( \frac{1}{w} \right)^2 \ast kp\_values \]

end

if \( \text{type} == \text{text} \) then

\[ \text{kaap} += \frac{1}{w} \ast kp\_values \]

end

end

\[ \text{kaap} = \frac{\text{kaap}}{\text{sum(kaap)}} \]

return \( \text{kaap} \)
Where the values of \( w \) weights: i) the sum of weights equals one; this is done to normalize the weights; ii) the weight of addition of \( f_i \) must be 34 times the weight \( w \) given by Eq. 4. The 'Marginal Contribution' of an edge connecting Node \( i \) and Node \( j \) is defined as the difference between the prediction probabilities on using their features. For a given predicted label \( c \), the marginal contribution of the feature \( f_i \) for the edge from Node 1 to Node 2 is calculated using Eq. 3. Here, \( prob\{f_i\} \) is the probability of label \( c \) calculated by having only feature \( f_i \) in the input and perturbing all other features to zero.

\[
MC_{f_i}\{f_i\} = prob\{f_i\} - prob\{\phi\} 
\]

(3)

To calculate the overall importance of \( f_i \), we need to calculate the weighted average of all 'marginal contribution' of \( f_i \) given by Eq. 4.

\[
KP_{\{f_i\}}(k) = w_{12} \times MC_{f_i}\{f_i\} + w_{34} \times MC_{f_i}\{f_i, f_2, \ldots, f_k\} 
\]

(4)

Where \( w_{12} \) and \( w_{34} \) are the weights for the weighted addition. Now, there are two conditions on the weights: i) the sum of weights equals one; this is done to normalize the weights; ii) the weight \( w_{34} \) must be \((k - 1)\) times the weight \( w_{12} \). The second condition is based on the fact that \( MC_{f_i}\{f_i\} \) is the effect of addition of \( f_i \) in an empty set of features, while \( MC_{f_i}\{f_i, f_2, \ldots, f_k\} \) is the effect of addition on a set containing \((k - 1)\) features. This results in Eq. 5.

\[
w_{12} + w_{34} = 1; \quad w_{12} = \frac{w_{34}}{k-1} 
\]

(5)

Where the values of \( w_{12} \) and \( w_{34} \), shown in Eq. 6 are computed using Eq. 5.
\[ w_{12} = \frac{1}{k}; \quad w_{34} = 1 - \frac{1}{k} \]  

(6)

The KP values shown in Eq. 7 are computed using Eq. 4 and Eq. 6.

\[ KP_{\{f_i\}}(k) = \frac{1}{k} \times MC_{f_i \{f_i\}} + \left(1 - \frac{1}{k}\right) \times MC_{f_i \{f_1, f_2, \ldots, f_k\}} \]  

(7)

3.3.2 Calculating KAAP values. This Section computes the KAAP values and uses them to determine the importance of each modality and its features. The information of image, text, and speech modalities are in the same data format, i.e., continuous format. A single pixel cannot define an object that can lead to a particular emotion for an image, but a group of pixels will. For speech, the spectrogram at a single instance of time & frequency alone can not define anything, but a time interval will. Likewise, for text, a single letter may not define an emotion, but a word is capable of doing so. KAAP values have been defined based on the motivation from the aforementioned fact. They are computed using the KP values for a group of features.

First, the input of size \( l \) is divided into \( k \) parts, where \( k \) is a hyperparameter decided through ablation study in Section 4.3. These \( k \) parts correspond to the \( k \) features of the input. Then, for a feature group \( f_i \), \( KP_{f_i}(k) \) values are computed for the given value of \( k \) using Eq 6. It represents how a group of features \( f_i \) will perform compared to all remaining groups. However, these groups can vary in size, i.e., \( k \) can have various values that lead to different groups and thus to different KP values from groups of different sizes, thus affecting the original features’ importance. To deal with this issue, the weighted average of all the KP values is taken for \( k \in \{2, 3, \ldots, k\} \) where weights are equal to the number of features in that group of features, given by the Eq. 8. It should be noted that \( k = 1 \) is ignored here, as the whole input as one feature will not make any sense.

\[ kAAP_{\{f_i\}} = \sum_{j=2}^{k} \left[ \frac{j}{l} \times KP_{\{f_i\}}(j) \right] \quad \text{for linear feature} \]  

(8)

For input image and speech spectrogram, both of width 128 and height 128, their KP values for a given \( k \) are calculated by dividing the input into \( k \) parts along both the axes. As a matrix defines both image and speech spectrogram, this gives us with \( k \times k \) feature group, the equation for calculating the KAAP values for the above two inputs is given by Eq. 9. It gives us a matrix showing the importance of each pixel for a given image and speech input. This matrix directly represents the importance of the image. At the same time, for speech input, the values are averaged along the frequency axis to reduce the KAAP value matrix to the time axis, hence giving importance to speech at a given time.

\[ kAAP_{\{f_i\}} = \sum_{j=2}^{k} \left[ \frac{j^2}{w^2} \times KP_{\{f_i\}}(j) \right] \quad \text{for matrix feature} \]  

(9)

For input text, the division is done such that each text word is considered a feature, as the emotion can only be defined by a word, not a single letter, as discussed above. Then the text is divided into \( k \) parts, and as a linear array can represent text, the KAAP values are calculated using Eq. 8. Also, the value of \( k \) used for image, speech, and text modalities have been determined as 7, 7, and 5 respectively in Section 4.3.2. Furthermore, the modalities’ importance defined by symbols \( \nu \), \( \delta \), and \( \tau \)
for visual, spoken, and textual features, respectively, are computed assuming that image, speech, and
text are three distinct features and calculating each modality’s KAAP value for $k = 3$. While finding
the importance of the features of a particular modality, all the other modalities are perturbed to zero.
The KAAP technique is depicted in Algorithms 3 which uses Algorithm 2 that calculates the KAAP
values for each data instance and 1 for probability prediction.

4 IMPLEMENTATION

4.1 Experimental setup

The network training for the proposed system has been carried out on Nvidia Quadro P5000 GPU,
whereas the testing & evaluation have been done on an Intel(R) Core(TM) i7-8700 Ubuntu machine
with 64-bit OS and 3.70 GHz, 16GB RAM.

4.2 Training strategy and hyperparameter setting

The model training has been performed using a batch-size of 64, train-test split of 80-20, Adam
optimizer, ReLU activation function with a learning rate of $1 \times 10^{-4}$ and ReduceLROnPlateau learning
rate scheduler with patience value of 2. The baselines and proposed models converged in terms of
validation loss in 10 to 15 epochs. As a safe upper bound, the models have been trained for 50 epochs
with EarlyStopping [52] with patience values of 5. The loss function used is the average of categorical
focal loss [32] and categorical cross-entropy loss. Accuracy, macro f1 [45], and CohenKappa [68]
have been analyzed for the model evaluation.

4.3 Ablation studies and models

The ablation studies have been performed to determine the threshold confidence value for data
construction, appropriate network configuration for VISTA Net, and suitable $k$ values for KAAP.

4.3.1 Ablation study 1: Determining baselines and proposed system’s architecture. To
begin with, the emotion recognition has been performed for a single modality at a time, i.e., separate
IER, SER, and TER using pre-trained VGG models [62] for Image & speech and BERT [12] for
text. The performance has been evaluated in terms of Accuracy, CohenKappa metric (CK), F1 score,
Precision, and Recall and summarized in Table 4. The CK metric measures if the distribution of the
predicted class is in line with the ground truth or not.

Table 4. Ablation Study 1. ‘Acc,’ ‘F1,’ ‘CK,’ ‘P,’ and ‘R’ denote accuracy, F1-score, CohenKappa score,
precision and recall.

| Model         | Acc   | F1   | CK   | P    | R    |
|---------------|-------|------|------|------|------|
| Image only    | 60.44 | 0.60 | 0.324| 0.60 | 0.60 |
| Speech only   | 78.69 | 0.75 | 0.624| 0.74 | 0.79 |
| Text only     | 81.51 | 0.81 | 0.69 | 0.81 | 0.82 |
| Image + Text  | 86.40 | 0.86 | 0.77 | 0.86 | 0.86 |
| Image + Speech| 84.66 | 0.85 | 0.746| 0.85 | 0.85 |
| Text + Speech | 81.95 | 0.81 | 0.70 | 0.82 | 0.81 |
| Image + Speech + Text | 86.60 | 0.86 | 0.78 | 0.86 | 0.87 |

Next, we moved on to the combination of two modalities. The chosen two modalities are fed
into respective pre-trained models and then passed from a dense layer of 512 neurons. Then the
information from these modalities is added using the WeightedAdd layer defined in 3.2.1, this output
is next passed from three dense layers of size 1024, 1024 and 4 neurons, which then classifies the
emotion. Image + text comes out to be the best combination, beating the remaining two combinations in both Accuracy and CK values.

At last, the information from all three modalities is combined and fed into their respective pre-trained models and is then passed from a dense layer of size 512, which is then passed from a WeightedAdd layer; the output of this layer is passed from 3 dense layers as in the combination of two modalities. Combining all three modalities has performed better than the remaining models in all the evaluation metrics. As observed during the experiments above, combining the information from the complementary modalities has led to better emotion recognition performance. Hence, the baselines and proposed model have been formulated, including all three modalities and various information fusion mechanisms in Section 4.4.

### 4.3.2 Ablation study 2: Determining $k$ values for KAAP

An in-depth ablation study has been conducted here to decide the value of $k$ used in Section 3.3.2. The dice coefficient [11] is used to determine the best $k$ values. It measures the similarity of two data samples; the value of 1 denotes that the two compared data samples are completely similar, whereas a value of 0 denotes their complete dis-similarity. For each modality, KAAP values are calculated at $k \in \{2, 3, \ldots, 10\}$. The dice coefficient is calculated for two adjacent $k$ values. For example, at $k = 3$, the KAAP values at $k = 2$ and $k = 3$ are used to calculate the dice coefficient. The procedure mentioned above has been performed for all three modalities, and the results are visualized in Fig. 6.

![Ablation Study 2: Determining appropriate $k$ values.](image)

The effect of increasing $k$ values can be observed in the figure. For image & speech, the value converges to 1 at $k = 7$, while for text, the optimal value of $k$ is 5.

### 4.4 Baselines and proposed models

The ‘Image + Speech + Text’ configuration described in Section 4.3.1 is considered as baseline 1, whereas further baselines models’ architectures have been formulated by incorporating further improvements in the information fusion mechanisms.

The baseline models are made on a common idea as described below. Firstly all the three modalities are fed into $P_i$, $P_s$, $S_i$, $S_s$, $T_i$ and $T_s$ as described in Section 3.2, and are then passed from a dense layer.
of 512 neurons, resulting in a 512-dimensional outputs which are then combined using WeightedAdd to give three outputs. The following strategy is being followed for combining them: any pre-trained network must be combined with another simpler network. At least one combination must contain the network from different modalities because if all the modalities combine with themselves, then such a combination will not lead to any information exchange. Thus, six such configurations are possible, as described in Eq. 10.

\[
\begin{align*}
(#1) & : \{P_i + S_i, P_s + S_s, P_t + S_t\} \\
(#2) & : \{P_i + S_i, P_s + S_t, P_t + S_s\} \\
(#3) & : \{P_i + S_s, P_s + S_t, P_t + S_t\} \\
(#4) & : \{P_i + S_s, P_s + S_t, P_t + S_i\} \\
(#5) & : \{P_i + S_t, P_s + S_s, P_t + S_s\} \\
(#6) & : \{P_i + S_t, P_s + S_s, P_t + S_i\}
\end{align*}
\]

The configuration (#1) is discarded as it does not hold the condition that at least one combination must combine with a different modality. The configurations (#2), (#3), (#6) are partially-complete combinations as one of the three outputs of these combinations combine the pre-trained and simpler network from the same modalities. On the other hand, the configurations (#4) and (#5) are complete.

Using the above strategy puts us in two disadvantages: i) only two out of five such baselines are complete while others are partially-complete; ii) different datasets have different requirements. For example, a particular multimodal dataset may have better images and speech components, while other datasets may have a better quality of text components. To generalize for any dataset and scenario, an automated multimodal emotion recognition system, VISTA Net, has been proposed, which combines all output of baselines 2-6 leaving any self combination and taking the weighted average of remaining all. Hence, it automatically decides the weights of each combination according to the requirements of problem statements and the dataset. The baselines’ and proposed system’s results are summarized in the following Section in Table 5.

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 VISTA Net has achieved emotion recognition accuracy of 95.99%. Its class-wise accuracies are shown in Fig. 7 while its results, along with the results of baselines, are shown in Table 5.

5.2 Qualitative results
Fig. 8 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, respectively. As observed from Fig. 8, speech and text were the most contributing modalities for the prediction of ‘angry’ and ‘hate’ classes, whereas image and text modalities contributed equally to the determination of ‘happy’ and ‘sad’ classes.

5.3 Results comparison
The emotion recognition results have been reported in Section 5.1. The IIT-R MMEmoRec dataset has been constructed from the B-T4SA dataset in this paper; hence, there are no existing emotion
Table 5. Results comparison for emotion recognition on IIT-R MMEmoRec dataset. ‘Acc,’ ‘F1,’ ‘CK,’ ‘P,’ & ‘R’ denote accuracy, F1-score, CohenKappa score, precision & recall. Baseline 1 has (Image + Speech + Text) configuration from Section 4.3.1.

| Model       | Acc   | F1    | CK    | P    | R    |
|-------------|-------|-------|-------|------|------|
| Baseline 1  | 86.60 | 0.86  | 0.78  | 0.86 | 0.87 |
| Baseline 2(#2) | 94.89 | 0.95  | 0.91  | 0.95 | 0.95 |
| Baseline 3(#3) | 95.44 | 0.95  | 0.92  | 0.93 | 0.95 |
| Baseline 4(#4) | 95.39 | 0.95  | 0.92  | 0.95 | 0.95 |
| Baseline 5(#5) | 95.58 | 0.96  | 0.92  | 0.96 | 0.96 |
| Baseline 6(#6) | 95.37 | 0.95  | 0.92  | 0.95 | 0.95 |
| VISTA Net   | 95.99 | 0.96  | 0.93  | 0.96 | 0.96 |

Fig. 8. Sample results; here, ‘P’, ‘GT’ are the predicted and ground-truth labels whereas and ‘Score’ denotes the importance of visual (υ), spoken (δ) & textual (τ) modalities.
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 have been compared with
VISTA Net’s sentiment classification results in Table 6.

Table 6. Results comparison for sentiment classification on BT4SA dataset with existing approaches. Here, ‘V,’ ‘S,’ and ‘T’ denote visual, spoken and textual modalities.

| Approach                              | Modality | Accuracy |
|---------------------------------------|----------|----------|
| Cross-Modal Learning [67]             | V + T    | 51.30%   |
| Multimodal Sentiment Analysis [16]    | V + T    | 60.42%   |
| Hybrid Fusion [29]                    | V + T    | 86.70%   |
| Automated ML [33]                     | V + T    | 95.19%   |
| VISTA Net (for Sentiment Classification) | V + S + T | 96.59%   |

5.4 Performance for missing modalities

In real-life scenarios, some of the data samples in the multimodal data may be missing information
about one of the modalities. The VISTA Net has been evaluated for such scenarios. We formulated
four use-cases with image, speech, text, or no modality missing respectively and divided the test
dataset into randomly selected equal parts accordingly. Then the information of the missing modality
has been overridden to null, and VISTA Net has been evaluated for emotion recognition. Table 7
summarizes the results thus observed.

Table 7. Results for missing modalities. Here, ‘Acc,’ ‘F1,’ ‘CK,’ ‘P,’ and ‘R’ denote accuracy, F1-score,
CohenKappa score, precision and recall.

| Model              | Acc  | F1  | CK  | P   | R   |
|--------------------|------|-----|-----|-----|-----|
| Missing Image      | 82.59| 0.82| 0.77| 0.86| 0.83|
| Missing Speech     | 57.62| 0.45| 0.75| 0.75| 0.58|
| Missing Text       | 62.82| 0.68| 0.70| 0.87| 0.63|
| Missing None       | 95.90| 0.96| 0.92| 0.96| 0.96|

As observed from Table 7, the emotion recognition performance for Missing no modality (i.e.,
having the information from all three modalities) is in line with the results observed in Section 5.2.
Further, missing image modality information has caused the least dip in the performance. Moreover,
the information from speech and text modalities combinedly has resulted in emotion classification
accuracy of 82.59%, whereas including all the modalities resulted in 95.90% accuracy. The afore-
mentioned observations are in-line with the observations in Section 4.3.1 where IER performance
was lesser than TER and SER performance.

5.5 Discussion

Various research tasks may require a particular modality’s information more than the others; for
example, text and visual information may be secondary for multimodal speech recognition. Likewise,
a multimodal emotion dataset might contain better quality information for a particular modality than
other modalities. In such cases, it would require human intervention to decide which modality is
more important for the analysis. However, the VISTA Net is capable of deciding that automatically. It
considers all possible combinations of various modalities’ information and weighs them accordingly.
As the proposed MMEmoRec dataset contains the information of complementary modalities, it enables the deep learning models to learn the contextually related representation of the underlying emotions. The ground truth is obtained by applying unimodal models. The final label is obtained by averaging the probability of each emotion obtained for each model and is considered the ground truth of the dataset. If the same unimodal emotion recognition models are used for emotion recognition during the dataset construction, then there will be a slight bias in the final performance. However, there will be no bias in developing and using a newer multimodal emotion recognition model. Furthermore, the human evaluators evaluated the IIT-R MMEmoRec dataset for the consistency of the determined emotion labels and appropriateness of the speech component, which has been synthesized via text-to-speech.

The qualitative & qualitative results (Fig. 7 & 8 and Table 5 & 6) have affirmed the importance of utilizing the information from complementary modalities. As observed from Fig. 8, different modalities have played a key role in determining the overall emotion portrayed by the input data sample. In some cases, the information for a particular modality may be missing from some of the data samples. The proposed system, VISTA Net, has been evaluated for such cases with missing modality information, and the observations are in accordance with the results previously observed and the insights gained during ablation studies.

The proposed interpretability technique, KAAP, computes the importance of each modality and the importance of their respective features towards the prediction of a particular emotion class. The existing interpretability techniques such as SHAP and LIME are not applicable to speech modalities, whereas KAAP is applicable to all image, text, and speech modalities. The proposed technique is expected to pave the way for growth in multimedia emotion analysis. We also hope that the IIT-R MMEmoRec dataset will inspire further advancements in this context.

6 CONCLUSIONS AND FUTURE WORK

The proposed system, VISTA Net, performs emotion recognition by considering the information from image, speech & text modalities. It combines the information from these modalities in a hybrid manner of intermediate and late fusion and determines their weights automatically. It has resulted in better performance on including image, speech & text modalities than including only one or two of these modalities. The proposed interpretability technique, KAAP, identifies each modality’s contribution and important features towards predicting a particular emotion class. The future research plan includes working on transforming emotional content from one modality to another. We will also work on controllable emotion generation, where the output contains the desired emotional tone.

REFERENCES

[1] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. SoundNet: Learning Sound Representations from Ulabeled Video. In Advances in neural information processing systems (NeurIPS), 892–900.
[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal Machine Learning: A Survey and Taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 41, 2 (2018), 423–443.
[3] Erdenebileg Batbaatar, Meijing Li, and Keun Ho Ryu. 2019. Semantic Emotion Neural Network for Emotion Recognition from Text. IEEE Access 7 (2019), 111866–111878.
[4] David A Bromiatowski et al. 2021. Psychological Foundations of Explainability and Interpretability in Artificial Intelligence. National Institute of Standards and Technology (NIST) (2021).
[5] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive Emotional dyadic MOtion CApture data. Language Resources and Evaluation 42, 4 (2008).
[6] Javier Castro et al. 2009. Polynomial Calculation of The Shapley Value based on Sampling. Elsevier Computers & Operations Research 36, 5 (2009), 1726–1730.
[7] Ze-Jing Chuang and Chung-Hsien Wu. 2004. Multi-modal emotion recognition from speech and text. International Journal of Computational Linguistics & Chinese Language Processing, Volume 9, Number 2, August 2004: Special
Issue on New Trends of Speech and Language Processing 9, 2 (2004), 45–62.

[8] Yucei Cimtay, Erhan Ekmeckicoglu, and Seyma Caglar-Özhan. 2020. Cross-subject multimodal emotion recognition based on hybrid fusion. IEEE Access 8 (2020), 168865–168878.

[9] Ciprian Adrian Corneanu, Marc Oliu Simón, Jeffrey F Cohn, and Sergio Escalera Guerrero. 2016. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. IEEE transactions on pattern analysis and machine intelligence 38, 8 (2016), 1548–1568.

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

[11] Ruoxi Deng, Chunhua Shen, Shengjun Liu, Huibing Wang, and Xinru Liu. 2018. Learning to Predict Crisp Boundaries. In The European Conference on Computer Vision (ECCV). 562–578.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Big Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805 (2018).

[13] Moaaz El Ayadi, Mohamed S Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classifications, and databases. Pattern Recognition 44, 3 (2011), 572–587.

[14] M. Beatrice Fazi. 2020. Beyond Human: Deep Learning, Explainability and Representation. Theory, Culture & Society (2020).

[15] Ruth Fong, Mandalena Patel, and Andrea Vedaldi. 2019. Understanding deep networks via extremal perturbations and smooth masks. In The IEEE/CVF International Conference on Computer Vision (ICCV). 2950–2958.

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

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

[18] Hatice Gunes, Bjorn Schuller, Maja Pantic, and Roddy Cowie. 2011. Emotion Representation, Analysis and Synthesis in Continuous Space: A Survey. In Face and Gesture. IEEE, 827–834.

[19] Mousannif Hajar et al. 2016. Using youtube comments for text-based emotion recognition. Procedia Computer Science 83 (2016), 292–299.

[20] Alan Hanjalic. 2006. Extracting Moods from Pictures and Sounds: Towards Personalized TV. IEEE Signal Processing Magazine vol. 23, no. 2, pp. 90–100 (2006).

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

[22] Yen-Hao Huang, Ssu-Rui Lee, Mau-Yun Ma, Yi-Hsin Chen, Ya-Wen Yu, and Yi-Shin Chen. 2019. EmotionX-IDEA: Emotion BERT—An Affectional Model for Conversation. arXiv preprint arXiv:1908.06264 (2019).

[23] Udit Jain, Karan Nathani, Nersisson Ruban, Alex Noel Joseph Raj, Zhemin Zhuang, and Vijayalakshmi GV Mahesh. 2018. Cubic SVM Classifier Based Feature Extraction and Emotion Detection from Speech Signals. In 2018 International Conference on Sensor Networks and Signal Processing (SNSP). IEEE, 386–391.

[24] Dhiraj Joshi, Ritendra Datta, Elena Fedorovskaya, Quang-Tuan Luong, James Z Wang, Jia Li, and Jiebo Luo. 2011. “Aesthetics and Emotions in Images”. IEEE Signal Processing Magazine vol. 28, 5 (2011), 1548–1568.

[25] Samira Ebrahimi Kahou, Xavier Bouthillier, Pascal Lamblin, Caglar Gulcehre, Vincent Michalski, Kishore Konda, Sébastien Jean, Pierre Froumenty, Yann Dauphin, Nicolas Boulangers-Lewandowski, et al. 2016. Emonets: Multimodal deep learning approaches for emotion recognition in video. Journal on Multimodal User Interfaces 10, 2 (2016), 99–111.

[26] ]kim2018building1 Hye-Rin Kim, Yeong-Seok Kim, Seon Joo Kim, and In-Kwon Lee. [n. d.]. Building Emotional Machines: Recognizing Image Emotions through Deep Neural Networks. IEEE Transactions on Multimedia (T-MM) 20, no. 11, pp. 2980–2992, (2018) (n. d.).

[27] Pieter-Jan Kindermans et al. 2019. The (Un)reliability of Saliency Methods. In Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. 267–280.

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

[29] Puneet Kumar, Vedanti Khokher, Yukti Gupta, and Balasubramanian Raman. 2021. Hybrid Fusion Based Approach for Multimodal Emotion Recognition with Insufficient Labeled Data. In The 28th IEEE International Conference on Image Processing (ICIP). 314–318.

[30] Chul Min Lee, Shrikantan S Narayanan, et al. 2005. Toward detecting emotions in spoken dialogs. IEEE transactions on speech and audio processing 13, 2 (2005), 293–303.

[31] Shan Li and Weihong Deng. 2018. Deep facial expression recognition: A survey. arXiv preprint arXiv:1804.08348 (2018).

[32] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for Dense Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision (ICCV), 2980–2988.
[33] Vasco Lopes, Antonio Gaspar, Luis A Alexandre, and João Cordeiro. 2021. An AutoML-based Approach to Multimodal Image Sentiment Analysis. In The International Joint Conference on Neural Networks (IJCNN). IEEE, 1–9.

[34] Jaime Lorenzo-Trueba, Roberto Barra-Chicote, Ruben San-Segundo, Javier Ferreiros, Junichi Yamagishi, and Juan M Montero. 2015. Emotion transcription through adaptation in HMM-based speech synthesis. Computer Speech & Language 34, 1 (2015), 292–307.

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

[36] Jana Machajdik and Allan Hanbury. pp. 83–92, (2010). “Affective Image Classification using Features Inspired by Psychology and Art Theory”. In Proceedings of the 18th ACM international conference on Multimedia (MM).

[37] Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. 2017. Deep learning-based document modeling for personality detection from text. IEEE Intelligent Systems 32, 2 (2017), 74–79.

[38] Navonil Majumder, Soujanya Poria, Devamanuy Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. In The 31st AAAI Conference on Artificial Intelligence (AAAI), Vol. 33. 6818–6825.

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

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

[41] Qirong Mao, Ming Dong, Zhengwei Huang, and Yongzhao Zhan. 2014. Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks. IEEE Transactions on Multimedia (TMM) 16, 8 (2014), 2203–2213.

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

[43] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang. 2017. Automatic speech emotion recognition using recurrent neural networks with local attention. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2227–2231.

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

[45] Juri Opitz and Sebastian Burst. 2019. Macro F1 and Macro F1. arXiv preprint arXiv:1911.03347 (2019).

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

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

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

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

[50] Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. 2016. Fusing audio, visual and textual clues for sentiment analysis of multimodal content. Neurocomputing 174 (2016), 50–59.

[51] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances. arXiv preprint arXiv:1905.02947 (2019).

[52] Lutz Prechelt. 1998. Early Stopping - But When? In Neural Networks: Tricks of the trade. Springer, 55–69.

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

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

[55] Jia Rong, Gang Li, and Yi-Ping Phoebe Chen. 2009. Acoustic feature selection for automatic emotion recognition from speech. Information processing & management 45, 3 (2009), 315–328.

[56] Klaus R Scherer and Harald G Wallbott. 1994. Evidence For Universality and Cultural Variation of Differential Emotion Response Patterning. Journal of Personality and Social Psychology 66, 2 (1994), 310.

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

[58] LS Shapley. 1953. A Value for n-person Games, Contributions to the Theory of Games II.
[59] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning Important Features Through Propagating Activation Differences. In International Conference on Machine Learning (ICML). 3145–3153.

[60] Kush Shrivastava, Shishir Kumar, and Deepak Kumar Jain. 2019. An Effective Approach for Emotion Detection in Multimedia Text Data using Sequence based Convolutional Neural Network. Multimedia Tools and Applications 78, 20 (2019), 29607–29639.

[61] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv preprint arXiv:1312.6034 (2013).

[62] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 (2014). Accessed on 30.06.2022.

[63] Shamane Siriwardhana, Andrew Reis, and Rivindu Weerasekera. 2020. Jointly Fine Tuning ‘BERT-Like’ Self Supervised Models to Improve Multimodal Speech Emotion Recognition. INTERSPEECH (2020), 3755–3759.

[64] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Bjorn Schuller, and Stefanos Zafeiriou. 2016. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In 2016 IEEE international Conference on Acoustics, Speech and Signal Processing (ICASSP), 5200–5204.

[65] Samarth Tripathi, Sarthak Tripathi, and Homayoon Beigi. 2018. Multimodal Emotion Recognition on IEMOCAP Dataset using Deep Learning. arXiv preprint arXiv:1804.05788 (2018).

[66] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Bjorn W Schuller, and Stefanos Zafeiriou. 2017. End-to-end multimodal emotion recognition using deep neural networks. IEEE Journal of Selected Topics in Signal Processing 11, 8 (2017), 1301–1309.

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

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

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

[70] Nan Xu, Wenji Mao, and Guandan Chen. 2018. A Co-memory Network for Multimodal Sentiment Analysis. In ACM International Conference on Research & Development in Information Retrieval (SIGIR).

[71] Ruiying Xu, Tao Chen, Yunqing Xia, Qiu Lin, Bin Liu, and Xuan Wang. 2015. Word Embedding Composition for Data Imbalances in Sentiment and Emotion Classification. Cognitive Computation 7, 2 (2015), 226–240.

[72] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. 2018. Multimodal Speech Emotion Recognition using Audio and Text. In IEEE Spoken Language Technology Workshop (SLT), 112–118.

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

[74] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018. Multi Attention Recurrent Network for Human Communication Comprehension. In The 32nd AAAI Conference on Artificial Intelligence (AAAI).

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