IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 24, 2021, accepted January 26, 2022, date of publication February 7, 2022, date of current version March 1, 2022.

*Digital Object Identifier 10.1109/ACCESS.2022.3149214*

# Adaptive Multimodal Emotion Detection Architecture for Social Robots

**JUANPABLO HEREDIA[1], EDMUNDO LOPES-SILVA[2], YUDITH CARDINALE[3,4], JOSE DIAZ-AMADO[2,3], IRVIN DONGO[3,5], WILFREDO GRATEROL[4], AND ANA AGUILERA[6]**

[1]Computer Science Department, Universidad Católica San Pablo, Arequipa 04001, Peru
[2]Department of Electrical Engineering, Instituto Federal da Bahia, Vitoria da Conquista 45078-300, Brazil
[3]Electrical and Electronics Engineering Department, Universidad Católica San Pablo, Arequipa 04001, Peru
[4]Departamento de Computación y T.I, Universidad Simón Bolívar, Caracas 1080, Venezuela
[5]University of Bordeaux, Estia Institute of Technology, Bordeaux, 64210 Bidart, France
[6]Escuela de Ingeniería Informática, Facultad de Ingeniería, Universidad de Valparaíso, Valparaíso 2340000, Chile

Corresponding author: Ana Aguilera (ana.aguilera@uv.cl)

This work was supported in part by the Fondo Nacional de Desarrollo CientíFico, Tecnológico y de Innovación Tecnológica—Fondecyt as Executing Entity of Concytec through the Project RUTAS—Robots for Urban Tourism Centers, Autonomous and Semantic-Based under Grant 01-2019-FONDECYT-BM-INC.INV; and in part by the Universidad de Valparaiso under Grant INICI UVA20993.

**ABSTRACT** Emotion recognition is a strategy for social robots used to implement better Human-Robot Interaction and model their social behaviour. Since human emotions can be expressed in different ways (e.g., face, gesture, voice), multimodal approaches are useful to support the recognition process. However, although there exist studies dealing with multimodal emotion recognition for social robots, they still present limitations in the fusion process, dropping their performance if one or more modalities are not present or if modalities have different qualities. This is a common situation in social robotics, due to the high variety of the sensory capacities of robots; hence, more flexible multimodal models are needed. In this context, we propose an adaptive and flexible emotion recognition architecture able to work with multiple sources and modalities of information and manage different levels of data quality and missing data, to lead robots to better understand the mood of people in a given environment and accordingly adapt their behaviour. Each modality is analyzed independently to then aggregate the partial results with a previous proposed fusion method, called EmbraceNet+, which is adapted and integrated to our proposed framework. We also present an extensive review of state-of-the-art studies dealing with fusion methods for multimodal emotion recognition approaches. We evaluate the performance of our proposed architecture by performing different tests in which several modalities are combined to classify emotions using four categories (i.e., happiness, neutral, sadness, and anger). Results reveal that our approach is able to adapt to the quality and presence of modalities. Furthermore, results obtained are validated and compared with other similar proposals, obtaining competitive performance with state-of-the-art models.

**INDEX TERMS** Emotion recognition, multimodal models, social robots, fusion process.

## I. INTRODUCTION

In people social interactions, emotion detection is a natural process that directly affects people's decision-making and actions during communication. In social robotics, this behaviour is mimicked by robots to interact naturally and harmoniously with people. To do so, robots can detect the emotion of human beings through visual perception [1], speech [2], nonverbal communication [3], mutual interaction [4], among others methods. In this sense, new proposals for social robots to detect emotions have become more naturalized and faster in recent years for better understanding of how to communicate with people [5]. In particular, proposals based on deep learning have increased with the perspective of more humanized applications in a more connected world, with tendencies to the seamless incorporation of robots into human environments [6]–[8].

Social robots catch the information needed to detect human emotions through their perception system. The main sources of perception in a robot are visual and audio sensory capacities [9]. Robots can get images and videos from their visual capacities and perceive speech with their audio

The associate editor coordinating the review of this manuscript and approving it for publication was Yi Zhang.

sensory capacity. Both sources, as well as other sources as touch sensing, can be combined for better identification of human emotion during the interaction, leading to a multisource approach. Emotion recognition is a strategy that makes robots able to assimilate human behavior in real social environments, especially for Human-Robot Interaction (HRI) [10].

The identification of human emotions through a sequence of images (video) allows defining a continuous and real process, helping to give greater veracity to the feeling found [11]–[13]. A video usually provides audiovisual information that can also be considered as different sources (e.g., audio, images, text), which in turns can be analyzed with different modalities. For example, from the audio, the content of the speech and voice modulation can be analyzed; from images, it is possible to analyze human faces, human postures, and the context to detect the emotion. Obviously, multisource methods are implicitly multimodal. Multisource and multimodal emotion detection approaches can capture the expression of the human's feeling in different ways, which can improve the classification process results in terms of precision and accuracy of the identified emotion [5], [14], [15]. To do so, a fusion method that aggregates the individual results is one of the most important aspects of multimodal approaches [16].

Which modalities are considered in multimodal emotion recognition methods depend on the purpose of the application and the available data. However, most multimodal approaches require having all modalities considered to show good performance, even though the presence of all of them is not guaranteed and their quality could be very uneven. Therefore, the fusion method must ensure that all available data are used properly, according to their quality and presence [16].

Although there exist studies dealing with multimodal emotion recognition for social robots [7], [17], [18], they still present a limitation in the fusion process: they can drop their performance if one or more modalities are not present or if modalities have different qualities. This is a common situation in social robotics, since robots can have a high variety of sensory capacities and might capture the word through different sources and with different levels of quality; hence, more flexible multimodal models are needed.

In this context, we propose an adaptive and flexible emotion recognition architecture able to work with multiple modalities and sources, as well as managing different levels of data quality and missing information, to lead robots to better understand the mood of human beings in a given environment to accordingly adapt their behaviour. Thus, the first contribution of this work is an efficient and adaptive architecture able to manage different inputs (e.g., videos, images, sound), as well as supporting missing and uneven quality of data, by integrating and adapting a previously proposed fusion method, called EmbracetNet+ [19]. Since each modality is processed independently, users can integrate any other modality, as well as substitute the provided modalities processing models. These characteristics make the framework appropriate to be applicable and adaptable for robots, who have different sensory capacities (different modalities) and might provide different levels of data quality. Nevertheless, it is still possible to model the robot behaviour depending on the detected emotion.

In the current version of our architecture, we consider that robots gather videos and speech through their sensory capacity. Videos are processed frame by frame by face emotion recognition models, and speech is transcribed to text, to then be analyzed with Natural Language Processing (NLP) techniques and analyzed with Mel Frequency Cepstral Coefficient (MFCC) to extract vocal frequency contrast. Our proposed architecture allows integrating other modalities, such as posture, context, touch sensing, by incorporating the corresponding individual modality processing model and adapting the EmbracetNet+ fusion method.

Another contribution of this work is an extensive review of state-of-the-art studies focused on fusion methods for multimodal emotion recognition approaches. Our revision demonstrates that existing approaches are not appropriate for social robots; thus, it is an obvious need of more flexible multimodal emotion detection models with adaptive fusion methods, in the context of social robots.

We evaluate the performance of our proposed architecture by performing different tests in which several modalities are combined to classify emotions using four categories: happiness, neutral, sadness, and anger. In total, we design four scenarios: (i) bimodal approach based on text and audio; (ii) bimodal approach based on text and face; (iii) bimodal approach based on face and audio; and (iv) a multimodal test considering text, audio, and face. Results reveal that our approach is able to adapt to the absence of modalities, but still reaching competitive performance compared with other multimodal proposals.

The remainder of this article is organized as follows. Section II shows some preliminary concepts of individual emotion detection and the classical methods used in face, text, and speech emotion detection approaches. Section III presents a review of relevant works related to fusion methods and multimodal emotion recognition studies for social robots. Section IV introduces the methodological approach adopted in this work, which focuses on a multimodal emotion detection approach to analyze videos taken from social robots. The experiments and results are shown in Section V. Section VI discusses improvements that can be carried out, as well as a summary of lessons learned. Finally, we draw conclusions in Section VII.

## II. PRELIMINARES

In general, multimodal emotion detection approaches analyze each modality separately; afterward, a fusion method is applied to aggregate the individual results. In this section, we present general descriptions of the most current popular solutions to detect emotions from faces, from text, and from speech, since they are the modalities considered in this work.

### A. FACIAL EMOTION DETECTION

This topic consists of recognising and classifying images of faces into emotions based on the detected facial features [20]. Those facial features are commonly the movement of facial muscles that build facial expressions [20]. Current facial emotion detection methods are based on Convolutional Neural Networks (CNN), whose inputs are images or videos [21]. Some of the CNN methods apply image feature extractor well-established methods for processing the facial data [22] that might be modified to become more suitable and achieve better results. Also, deep learning approaches based on Attentional CNN are becoming useful for detect emotions from faces, where only more relevant face parts (facial features) in the image are explored [23].

### B. TEXT EMOTION DETECTION

The field of NLP, despite not being new, has had mayor advancements in the past decade. The mainstreaming of pretrained Word Embeddings allows a large portion of the words in a language to be represented in a vector space that maintains a spatial relation between words based on their semantic meaning. Due to the sequential nature of sentences, one of the first widely adopted architectures for text classification is Recurrent Neural Networks (RNN), but they were later replaced by some variants of them, such as the Gated Recurrent Units (GRU) [24]–[26], Long Short-Term Memory (LSTM) [27], and multiplicative LSTM (mLSTM) [28]. Given that most emotion classification task in text can be modeled as a text classification task, these approaches and models have all been used to solve this type of problems as well.

Usually, these models are not perfect and require a lot of task specific labelled data for the training process, which it is hard to found or obtain. Thus, some other methods are needed to overcome this limitation, such as the Attention-based Transformer architecture, that posses a great potential for transfer learning. It can be trained with unlabelled data, with unsupervised learning process, and then fine-tuned for downstream task with a relatively small amount of labelled data compared to other models.[1] There are currently different available implementations of transformers publicly available through libraries, such as the *Hugging Face Transformers* library [29], which is at the core of many other libraries.

### C. SPEECH EMOTION DETECTION

The voice is a primary resource that is widely used and extremely important for communication among human beings; it transmits "more than simple words''. Thus, the use of speech to explore emotions is significant due the rich information that can be extracted. To perform the recognition of emotions through speech, it is first necessary to extract and classify the resources. Thus, it is common to use resource extraction processes in audio files, such as the Fourier parameters [30] obtained from the Discrete Fourier Transform. After extracting this data, resource classifications must be made using linear and non-linear classifiers. The most commonly used linear classifier is the Support Vector Machine (SVM), but nonlinear classifier models guarantee better effective results for low-level activities, including Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM). Mel Frequency Cepstral Coefficient (MFCC) and Linear Predictor Coefficients (LPC) are more used in Speech Emotion Recognition (SER) models and with the computational power of deep learning, there are greater capacities to automatically detect complex structures and resources. Vocal frequency contrast can differentiate an emotional state and be perceptible by neural networks trained for emotion recognition, there exist several approaches to implement SER models, for example, based on CNN combined with Random Forest and extraction of speech emotion features from the normalized spectogram [31] or based on deep neural networks and extreme machine learning [32], [33].

## III. RELATED WORK

More related studies to our work are those focusing on multimodal emotion detection, including facial, audio, and text, as well as on significant fusion methods. We survey in this section some relevant works in these regards. In addition, the implementation and association with emotion detection in social robots are also highlighted.

### A. FUSION METHODS IN MULTIMODAL EMOTION DETECTION MODELS

Multimodal emotion detection methods can be classified into early fusion methods and late fusion methods [43]. Early fusion consists in simply concatenating multimodal features mostly at input level [44]. Methods that fall in this group, perform the fusion on the input data and then process the fusion result as a single complex data. These methods are disused because the difference and heterogeneity of the input do not allow exploiting correctly every source. On the other hand, in late fusion methods, the data are merged at the end – i.e., data of each source are processed separately and then the results are merged. These methods are the most used and they have presented state-of-the-art results. In this section, we review some relevant and recent late fusion approaches, particularly those that apply machine learning methods to merge and obtain an estimated emotion.

We review two groups of late fusion methods: those based on Multi Layer Perceptron (MLP) [33]–[35] and those based on more complex models, such as combinations of CNN, RNN, LSTM, and others [36]–[39]. Table 1 summarizes these studies. In [34], authors propose a complete system to recognise emotions from facial gestures and audio-text features. Their system ends with a Multiplicative Fusion [54] and its integration into an MLP that outputs the corresponding emotion. In a more simple way, in [33] a source results concatenation is carried out, to then use an MLP that ends with a softmax function. This model processes facial, text,

[1] https://jalammar.github.io/illustrated-bert/ (accessed on February 26 2021)

**TABLE 1.** Emotion detection fusion methods.

| Work | Modality data | Dataset | Single Analysis Methods | Fusion Methods |
|---|---|---|---|---|
| Mittal et al. (2020) [34] | Face, text, and speech | IEMOCAP and CMU-MOSEI {happy, sad, angry, neutral} | MLP | MLP with multiplicative fusion |
| Tripathi et al. (2018) [33] | Speech, text, and motions | IEMOCAP {anger, excitement (happiness), neutral and sadness} | LSTM, CNN, MLP with Adam optimizers, word-embedding models and RNN decoders | MLP with a softmax function |
| Ortega et al. (2019) [35] | Face, speech and textual information | EmotiW {arousal, valence and Liking} | MLP | MLP with a single linear activation function |
| Poria et al. (2016) [36] | Video, audio, and text | Multimodal opinion utterances dataset (MOUD), YouTube videos of product reviews, ICT-MMMO for sentiments {positive, negative} and IEMOCAP for emotions {happy, sad, angry, and neutral} | CNN and RNN | Multiple kernel learning |
| Kahou et al. (2016) [37] | Audio and facial images in videos | EmotiW{angry, disgust, fear, happy, neutral, sad, surprise} | CNN, DBN, Autoencoders, shallow network | SVM and MLP Aggregation Techniques |
| Tzirakis et al. (2021) [38] | Audio, text and visual data | SEWA {arousal and valence} | CNN | An attention layer and 1-layer LSTM network |
| Akhtar et al. (2019) [39] | Text, acoustic and visual frames | CMU-MOSEI Sentiment {positive, negative} and emotion {anger, disgust, fear, happy, sad or surprise} | biGRU network | biGRU network with the outputs of a pair-wise attention mechanism concatenated |
| Sun et al. (2016) [40] | Audio and visual frames | Chinese Natural Audio-Visual Emotion Database (CHEAVD) and {anger, disgust, worried, sad happy, anxious, surprise, neutral} | SVM, MLP, LSTM, DCNN | Aggregate fusion with a majority vote algorithm and a weight shared fusion network |
| Heredia et al. (2021) [19] | Face, body, posture and context | EMOTIC (modified by grouping) {anger, anticipation, disgust, fear, joy, sadness, surprise, and trust} | CNN, DGCN | EmbraceNet+ |
| Lan et al. (2020) [41] | EEG, eye image (EIG), and eye movement (EYE) | SEED-V dataset {happy, sad, fear, disgust, and neutral} | PSD, DE, STFT ResNet, LSTM | DGCCA (Adaptive) |
| Li et al. (2019) [42] | Magnetoencephalogram (MEG) and Explicit and implicit emotional responses | DECAF Arousal: 0 (very calm) to 4 (very excited) and Valence: -2(very unpleasant) to 2 (very pleasant) | DCT, IBI, HR, HRV GRAD channels, PSD | HMNN (Adaptive) |
| **Our work** | **Face, speech, and text** | **IEMOCAP {anger, happiness, sadness, neutral}** | **CNN, RNN (transformer) can be changed, others can be added (adaptive)** | **EmbraceNet+ (Adaptive)** |

and speech data. In [35], authors also process audio (speech), visual (face), and text data. They process the data separately by MLP; then, the outputs are concatenated, it is the input for their fusion system that consists in another MLP, a single linear neuron and a scaling module, the final prediction is produced by a linear activation function.

On the more complex side, the study presented in [36], combines the modality of video, audio, and text to predict people feelings. Each kind of data is processed separately and each one gives one score, which are concatenated and got into a Multiple Kernel Learning (MKL). This fusion method is a feature selection method where features are organized into groups and each group has its core function; with the MKL, the state-of-art achieved by using a SVM was beaten. Authors in [37] also beat the state-of-art with SVM, for processing audio and facial images in videos. Instead, they experiment with several improvements over the SVM method. In [38], a method that uses audio, text, and visual data is presented. Data of each source are processed separately and then merged with a model that has an attention layer and a 1-layer LSTM network. The attention layer merges the modalities by concatenating them and using a residual approach to generate the fusion.

In [39], authors propose a deep multitask framework that uses both multimodal sentiment and emotion analysis. Their approach learns a joint-representation for both the tasks as an application of GRU based on an inter-modal attention framework.

In [40], it is explored a multi-feature based classification framework for the Multimodal Emotion Recognition Challenge with eight facial emotions in short video segments extracted from Chinese films, TV plays, and talk shows. SVM and Deep Convolutional Neural Network methods are used to extract various features. A decision-level fusion method is explored to aggregate the different prediction results using a majority vote algorithm and a weight shared fusion network. In [19], authors use features from context, face, posture, and body to recognize emotions. The fusion method is EmbraceNet+, an architecture composed by three fusion methods (EmbraceNet [45], weighed sum, and concatenation). This combined method improves learning of the correlation of modalities or among modalities and the results achieved. EmbraceNet is a trainable method based on the multinomial distribution. Both EmbraceNet and EmbraceNet+ are tolerant for missing modality data – i.e., data loss of a modality can be covered by the other modalities, and they could be applied in any machine/deep learning system.

Concerning adaptive multimodal emotion recognition models, where the fusion weights are adapted (giving less or more impact to the modalities) in order to increase the accuracy of results, there exist few works. Authors in [41] use stochastic gradient descent and adopt backpropagation to update the weights matrices. By using the SEED-V dataset,[2] which provides not only electroencephalography (EEG) signals but also eye movement features recorded by SensoMotoric Instrument (eye-tracking glasses[3]), authors obtained an accuracy of 82.11% and standard deviation of 2.76%. In another work [42], authors affirm that "feature-level fusion methods cannot deal with missing or corrupted data, while decision-level fusion methods may lose the correlation information between different modalities". For that, a Hierarchy Modular Neural Network (HMNN) is proposed. HMNN recognizes the result according to the activity level of each module (each class is a module), and the one with the maximal activity wins the competition, which is called the winner-take-all strategy. Experiments using DECAF dataset,[4] showed an improvement with respect to the state-of-the-art and its effectiveness in dealing with missing or corrupted data.

In the literature, there are fusion methods that are applied in different domains and use other types of sources (Table 2). For instance, the document classification, based on text and visual data [47]; in a medical context, for early detection of Alzheimer's disease stage using magnetic resonance imaging, genetic data, and clinical test data [48]; in music videos [49], [50]; or for multimodal languages [51]–[53].

Even though the focus of these works have not been the emotion recognition, they propose interesting network architectures with additional layers to combine multi modalities and outperforms the single modalities. In [47], authors introduce an end-to-end learnable multimodal deep network that jointly learned text and image features and perform the final classification based on a fused heterogeneous representation of the document. A concatenation strategy was used to combine both feature vectors into one. In [48], the proposal integrates the intermediate features generated for each modality using a concatenation layer followed by a classification layer to predict the Alzheimer stage. Authors try k-nearest neighbors, decision trees, random forests, and support vectors machines as alternatives for the classification layer. The missing modalities are masked with zeros.

An interesting point in the study presented in [49] is the network architecture, where learned information is integrated after each block of the dense residual network by using Multimodal Transfer Module (MMTM). Afterward, the MMTM and SE (squeeze and excitation) networks are used for information sharing and boosting during training. Earlier the same authors built their own music video dataset and demonstrated that the multimodal results show improvement in various evaluation matrices over unimodal performance [50]. The proposal presented in [51] is focused on the crossmodal attention mechanism, which provides a latent crossmodal adaptation that fuses multimodal information by directly attending to low-level features in other modalities.

In [52], authors model expressive nonverbal representations by analyzing the fine-grained visual and acoustic patterns that occur during word segments. Also, they propose a Recurrent Attended Variation Embedding Network to model the fine-grained structure of nonverbal subword sequences and dynamically shifts word representations based on nonverbal cues. In [53], it is presented a framework for fine-tuning BERT and XLNet for multimodal input. This framework allows the BERT and XLNet core structures to remain intact and only attaches a Multimodal Adaptation Gate to the models. Likewise, other fusion methods generalize to any approach, such as EmbraceNet [45], which is based on multinomial probabilistic theory to fuse feature vectors and CentralNet [46], where the fusion of the unimodal representations is done using a learned weighted sum. Multiple combinations are done at level of feature through direct concatenation using shallow models, intermediate using deep models, and decision by voting on the single-modalities using shallow models.

As shown in Table 1 and Table 2, the most of reviewed methods are trainable approaches based on MLP [33]–[35], [37], [47] and others are methods with more complex architectures [19], [36], [38], [39], [45], [46], [48]. However, the most of them are not automatically adaptable or suitable to work with wrong or missing data except [19], [45].

[2] SEED-V dataset is available in https://bcmi.sjtu.edu.cn/home/seed/index.html.

[3] The SMI eye-tracking glasses website - https://imotions.com/hardware/smi-eye-tracking-glasses/.

[4] DECAF dataset - http://mhug.disi.unitn.it/wp-content/DECAF/DECAF.html#/datasets

Even though some methods could be adapted for this, their performance could be affected due that they were trained without considering uncontrolled scenarios. Instead, the method proposed by [19], [45] was trained simulating loss or problems with data in order to obtain a more robust model and applicable to social robots, whose sensory capacities vary and might catch the world with different levels of qualities. In this work, we adopt the EmbraceNet+ fusion method proposed in [19] to manage such uneven inputs. EmbraceNet+ tolerates missing data by multiplying the probabilities of the multinomial distribution with a binary vector of availability. This capacity permits effectively an adaptive and flexible emotion recognition architecture with different modalities ideal for the social robot contexts. As it is shown in Table 1, our approach is similar to several works considering the multimodalities and dataset used, but differ from all others in the single analysis methods and the fusion method. Actuality, in our proposal, the single methods can be changed and others can be added, since the EmbraceNet+ can manage it.

### B. EMOTION RECOGNITION IN SOCIAL ROBOTS

Even though emotions have been explored in the context of HRI, it is currently still a challenge in this area. This is due to the necessity of reliable results, to provide a trustworthy interaction, and the time constraints required to account for the recognized emotion into the adaptation of the robot behavior [10]. The survey presented in [55] shows researches during the years 2000–2020 focusing on humans' recognition and responses to artificial emotions of social robots. The review advances in robotic psychology by revealing insights about the generation of artificial robotic emotions (stimulus), human recognition of robotic artificial emotions (organism), and human responses to robotic emotions (response). Other survey presented in [56] focuses on 232 papers categorized as considering emotional intelligence, emotional model, or implementation of the model. Results of this review demonstrate the utility of emotion recognition in robotics, mainly to improve HRI or to improve robots' performance.

Actually, many existing works in social robotics deal with face emotion recognition to improve HRI and social navigation. The study described in [57] presents a survey of 101 papers between 2000 and 2020 on facial emotion detection. This survey emphasizes the recognition of human facial expressions and the generation of robotic facial expressions. Authors declare that the accuracy on facial expression recognition in the wild is considerably lower than the experiments which have been conducted under controlled laboratory conditions. In the context of navigation algorithm for robots, authors of [58] propose a mood detector component. This component takes the image as input and generates a list of emotions detected in the faces existing in the image. The mood is persistent in time, and each person is tagged with a positive, neutral, or negative mood.

Few works adopt multimodal approaches in social robotics. In [7], authors work on a multimodal model for emotion detection from facial and speech with Two-Stage Fuzzy Fusion based CNN. After training the network, they use a camera and a kinect to capture data under test and perform a simulation to verify the accuracy of results achieved. Thus, there is an application in NAO Robot with the images captured by kinect and processed by the proposed Softmax-regression-based method and producing the output on a screen. The application of this system in robotics for real-time emotion detection produces a great revolution for the development of robotics in a more natural coexistence between human-robots and a better coordination of the performance of actions in daily activities/services in a social environment.

Authors in [18] report that limited work is presented in literature, wherein multiple modalities are considered for emotion recognition. In [17], authors present an interactive robot learning framework using multimodal data from thermal facial images and human gait data for online emotion recognition. Their decision-level fusion method for the multimodal classification is implemented using Random Forest model. The emotion recognition model is focused on the detection of four human emotions (i.e., neutral, happiness, angry, and sadness). In [18], a model is trained on facial and speech samples using 1-Dimensional and 2-Dimensional CNN and also trained using pretrained networks such as Visual Geometry Group-16 and Inception Version 3. A simple CNN network from facial emotion recognition and CNN network with log-Mel spectrogram as input from speech emotion recognition are used for the combined emotion recognition model. An interesting result is the difference of performance when the speech network is trained on male and female speakers' samples separately, the obtained accuracy is better compared to the combined dataset. Results in that work show that females are more expressive than male and combining image and speech modalities improves the recognition rate.

In [59], a robot Pepper is used to improve HRI with users with physical disabilities, based on the recognition of emotions from videos and speech. In [60], it is proposed a model able to recognize emotions from text and store this information in a semantic repository, based on an ontology to represent emotions. It is considered as a multimodal approach since text can be obtained directly as text or by converting speech in text.

Without pretending to be an exhaustive review, these studies reveal some limitations and some challenges are still open in the area of HRI. The combination of more sources and more modalities, and even further, the possibility of using deficiencies in emotion recognition from some sources or modalities over others are obvious challenges in this area. The improvement of the recognition performance in some single modalities, such as the speech and human movements and, the fusion of multiple modalities is necessary. To overcome some of these limitations, we propose an architecture which integrates popular processing models to analyze independently each modality to then combine the individual

**TABLE 2.** Fusion methods in other contexts.

| Work | Modality data | Dataset | Single Analysis Method | Context | Fusion Method |
|---|---|---|---|---|---|
| Choi and Lee (2019) [45] | Multimodal Imagen | The gas sensor arrays and OPPORTUNITY | Any DNN | Chemical sources and human activities | EmbraceNet |
| Vielzeuf et al. (2018) [46] | Multimodal Imagen | Multimodal MNIST, Audiovisual MNIST, Montalbano and MM-IMDb | Any DNN | Audiovisual and Sign gesture | Neural network and learned weighted sum |
| Audebert et al. (2019) [47] | Image and text | ImageNet, RVL-CDIP and Tobacco3482 | MLP, CNN | Document, classification | MLP with concatenation of individual output vectors |
| Venugopalan et al. (2021) [48] | Image and text | ADNI | Autoencoders, CNN | Medical data | Learned weighted sum |
| Pandeya et al. (2021) [49] | Music, video, and facial | Self-constructed | CNN | Music videos | Softmax + aggregated and computed class probability |
| Pandeya and Lee (2021) [50] | Music, video | Self-constructed | CNN | Music videos | Softmax decision operator |
| Tsai et al. (2019) [51] | Image, text and audio | CMU-MOSI MOSEI, IEMOCAP | Temporal Convolutions, Positional Embedding | Multimodal language (aligning the data) | Multimodal Transformer |
| Wang et al. (2018) [52] | Image, text and audio | CMU-MOSI, IEMOCAP | LSTMs | Multimodal language (word shifting) | Gated Modality-mixing Network and Multimodal Shifting |
| Rahman et al. (2021) [53] | Image, text and audio | CMU-MOSI MOSEI | BERT and XLNet | Multimodal language (finetuning large pre-trained Transformer models) | Multimodal Adaptation Gate (aggregation) |

results with an EmbracetNet-based fusion method, which is adaptable for missing data or different levels of quality. The general description of our proposal is presented in the following section.

## IV. MULTIMODAL SYSTEM TO RECOGNISE EMOTIONS

In social robotics, emotion recognition is commonly used to improve the making decision process of the robot. Our system is aimed at improving existing studies by considering the analysis of several modalities to improve the results while keeping its applicability for social robots and using their available sensors. This section describes the different components of the proposed system, including the capturing and pre-processing strategy of inputs, the individual processing of each modality, the fusion method able to manage missing data and uneven quality of data, and modeling the robot behaviour depending on the detected emotion. We also show how our framework is flexible and adaptive at every component in the context of social robotics.

The general pipeline of our current version of the proposed system is shown in Figure 1, in which each gear symbol means input processing, such as face recognition, audio transcription, and MFCC feature extraction. Video frames are processed at the time of their capture (Figure 1(a)); while speech data are transformed into text, by specialized software tools, and their MFCC features, relevant for the emotion detection, are extracted by using a non-linear scale (Figure 1(b)).

Afterwards, the text and audio modalities are processed to then merge them with the face modality from videos, with the EmbraceNet+ method (Figure 1(c)). Finally, the robot applies a procedure to change its behaviour according to the recognised emotion (Figure 1(d)). In the following, we describe each component in detail.

### A. MULTIMEDIA CAPTURING AND PROCESSING

As a multimodal approach, our proposal can consider videos, images, audio, text, and others as input data, captured by appropriate sensors of the sensory capacity of social robots (e.g., camera, microphone, touch screen). For example, images and videos can be captured with built-in cameras and audio through microphones, while text can be obtained directly through a keyboard, touch screen, or indirectly by translating speech into text. These inputs are gathered from people speaking among them or people interacting with the social robot.

The performance of the system depends on the quality of captured data and the sensors used. In the current version, audio data is recorded and processed in its entirety, unlike image data, which is processed at the time of capture, and then used the results of each frame together (for example, on average); this approach might reduce the execution time a bit. For each captured frame, the face region is recognized and processed separately; any method for facial recognition is applicable. Audio is processed in two ways: an MFCC features extractor is used to obtain a ‘‘graphic’’ representation of speech, and conversion into text or transcription of speech to be analyzed with NLP techniques. Speech-to-text conversion is an important technique and necessary for the development of a complete emotion recognition system [61] because the transcription of audio enables independent textual processing and analysis. If the recorded audio is

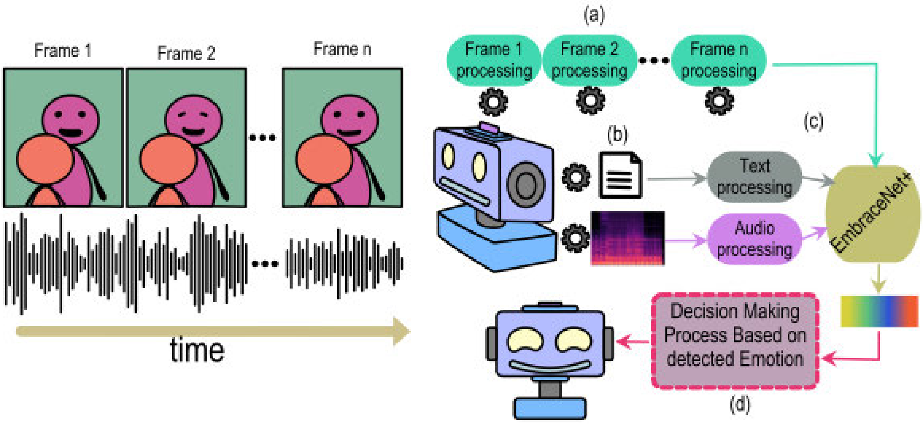


**FIGURE 1.** The overall pipeline of the proposed system. Each gear symbol means input processing such as face recognition, audio transcription, and MFCC feature extraction: (a) video frames are processed at the time of their capture; (b) speech data is transformed into text and MFCC features; (c) integration mechanism, the text and audio modalities are processed and then merge the three modalities with the EmbraceNet+ method; (d) procedure to change the robot behaviour according to the recognised emotion.

unclear, both text and audio are ignored and considered as unavailable data.

Moreover, whether the robot has more sensors (e.g., touchscreen, proximity lasers) or more special processing modules (e.g., body posture detector, gaze tracker, sign language translator), more data can be considered, and the system might have more modalities. While with more data to process the system will be heavier, it also will be more robust, inclusive, and malleable.

## B. INDIVIDUAL EMOTION RECOGNITION

To process each modality, the proposed system is aimed at using popular models of emotion detection. In the following, we explain the ones used in the current version of our architecture, as examples of possible models that can be integrated.

### 1) FACE MODALITY

The approach of Parkhi *et al.* [22] is taken, but we use a VGG19 architecture as the basis, instead of the VGG16 as proposed in [22]. In addition, the size of images is reduced to 48 × 48, which also reduce the number of extracted features from convolutional layers; thus, the linear layers at the end of the network also have fewer parameters. This VGG19 model has five blocks composed of convolutional layers and max-pooling layers, as shown in Figure 2. Those convolutional blocks have the following configuration: (i) two convolutions of 64 filters; (ii) two convolutions of 128 filters; (iii) four convolutions of 256 filters; (iv) four convolutions of 512 filters; and (v) four convolutions of 512 filters. After the convolution blocks, an average-pooling layer is applied, resulting in a vector of 512 features after flattening the results of the convolutions. Finally, a linear layer processes the 512 features to get the final prediction.

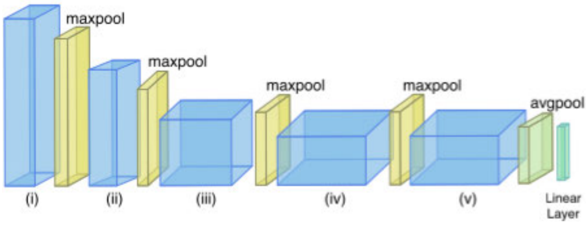


**FIGURE 2.** Architecture of VGG-face model.

This approach works effectively and in lower time than VGG architectures, because of the shorter input size and fewer features to process, which is desirable for the embedded hardware of social robots.

### 2) AUDIO MODALITY

For the emotion detection from speech, a convolutional model used by Venkataramanan and Rajamohan [62] is applied (see Figure 3). For this model, the voice audios are processed as MFCC and used as 2D data; and since they are 2D vectors, they could be processed as images to resize them and make the input data uniform. This input size is set to 128 × 259, which means 128 audio features and 259 units of time. The architecture of the model has four blocks, where each has a convolution layer, a batch normalization, an activation function, a max-pooling layer, and a dropout layer. The convolution layers are configured as L1:128, L2:128, L3:64, L4:64; the pooling layers as L1:2 × 2; and the others as L2, L3, L4: 4 × 4. The dropout layers are set as 0.5 for the first block (L1) and 0.25 for the others (L2, L3, L4). Finally, the activation function used is the Exponential Linear Unit (ELU) function, which allows some negative values and helps the learning process; however, the execution time could increase due to the added exponential operation. Nevertheless, during model training, ELU allows faster convergence than ReLU

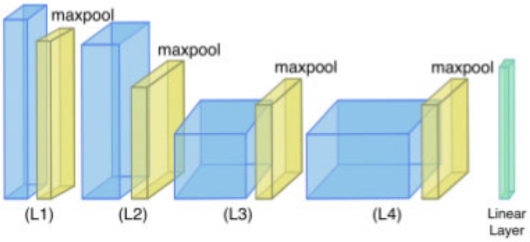


**FIGURE 3.** Architecture of the audio model.

and its variants; but during testing, the model with ELU will run slower.[5]

### 3) TEXT MODALITY

For this modality, we use a PyTorch implementation of DialogXL[6] [63], which is a model specialized for Emotion Recognition in Conversation (ERC) based on XLNet [64]. This model has been compared against the Conversational Memory Network (CMN), DialogueRNN, HiGRU, DialogueGCN, TL-ERC, KET, BERT, and XLNet models, as well as a stacked ensemble for emotion recognition [60], and it has shown the best performance out of all of them. The original version of XLNet is a model that has reached state-of-the-art results in many applications. It consist of an auto-regressive language model based on the transformer architecture which, given a sequence, outputs the probability of the sequence of words to follow.[7] DialogXL improves on XLNet by using an enhanced memory to store longer historical context during dialog and Dialog-Aware Self-Attention to keep track of the different speakers in a conversation.

Each sentence by a speaker (utterance) is passed through an embedding layer which tokenizes the sentence into a sequence of vectors as shown in Figure 4. This representation is then passed through a stack of neural networks where each layer contains a Dialog-Aware Self-Attention component and an Utterance Recurrence Component; each of these layers output a vector that is passed to the following layer. At the end of the last layer, the hidden state of the classification token and the historical context are passed through a feed-forward neural network to get the predicted emotion.

### 4) OTHER MODALITIES

Adding more modalities include the addition of a capturing module and a modality processing module. For instance, for recognizing emotions from body postures and hand gestures, some deep learning methods of human pose detection can be used to capture the data; and to process those data, a new trend is to use graph neural networks, especially the convolutional ones that extract pertinent features [19]. On the other hand, joining sensors like keyboards and touchscreens, or even a sign language translator, open doors to better analyze the emotions of a deaf-mute person. Although the embedding data captured could be processed similarly to our text modality; thus, this would not be completely a new modality but a relevant complement for the text one.

## C. FUSION METHOD

In this study, the EmbraceNet+ method [19] is used to merge the modalities and generate the final prediction. EmbraceNet+ is an extension of EmbraceNet [45] that improves some aspects, and it is composed of three EmbraceNets and two additional fusion methods. A basic EmbraceNet has docking layers that adjust the modality outputs to the same size (*embracement_size*) while learning about the correlation between them; and an embrace layer that selects features from the modalities outputs following the multinomial distribution. Unlike, in EmbraceNet+ two basic EmbraceNets have modifications in docking layers by adding a linear layer and a dropout layer in each docking layer. There are as many docking layers as modalities and only one embrace layer in the basic EmbraceNet, even though the embrace layer needs as many probabilities and values of availability as modalities to perform correctly. In the proposed system, the individual models used only give the final classification, not the vectors of intermediate characteristics; therefore, in the EmbraceNet+ fusion method, the initial EmbraceNet, which processes the intermediate data, is removed (see Figure 5).

Concretely, each altered docking layers are made up of a linear layer of 32 neurons ($D_{1,1}$), a dropout layer with 0.5 of decay probability, and another linear layer of 16 neurons ($D_{1,2}$). As additional fusion methods, we use the weighted sum, whose output is a vector of $n$ probabilities ($n =$ number of emotion categories), and a concatenation, whose output is a vector of $3n$ because of the number of modalities. Thus, the other EmbraceNet receives three vectors of 16, $n$, and $3n$ values (that work as modalities), and are handled by docking layers of one linear layer of 16 neurons each one ($d^{(k)}$), which conducts to add an extra linear layer of $n$ neurons, which outputs the final prediction. Inside each EmbraceNet, the received data pass through a corresponding docking layer (Alg. 1 line 2) and form a vector of *embracement_size* × *N_modalities* that is compacted at the embrace layer phase.

The EmbraceNet and the EmbraceNet+ tolerate missing data by multiplying the probabilities of the multinomial distribution with a binary vector of availability (Alg. 1 line 4). This availability vector contains 1 if the respective modality data is correct, otherwise, it contains 0. The method also assures that $\sum_i p_i = 1$, where $i$ indicate the modality, and $p$ the corresponding probability for each modality (Alg. 1 line 5). For instance, if the face data is unclear or erroneous, the availability vector will be [0, 1, 1] and the probabilities [0.0, 0.5, 0.5]; assuming the vector represents [*face*, *audio*, *text*]. Those probabilities are used to select features following the multinomial distribution (Alg. 1 line 6),

[5] https://deeplearninguniversity.com/elu-as-an-activation-function-in-neural-networks/

[6] https://github.com/shenwzh3/DialogXL

[7] https://www.borealisai.com/en/blog/understanding-xlnet/

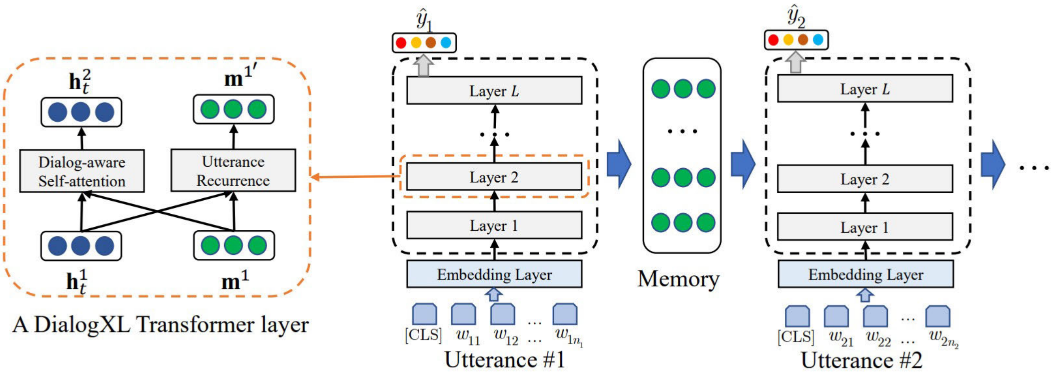


**FIGURE 4.** The architecture of DialogXL as proposed by the original authors [63].

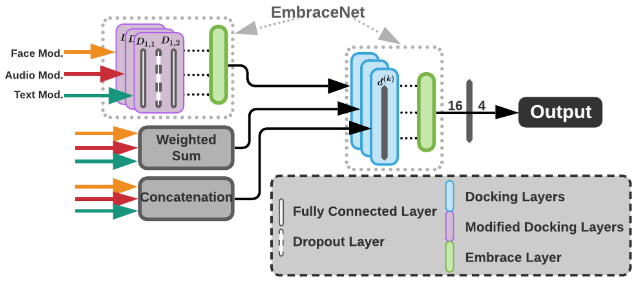


**FIGURE 5.** Adapted EmbraceNet+ architecture..

**Algorithm 1** EmbraceNet

```
Require: Modalities predictions array, availability vector,
    and the probabilities of the modalities
docking_output ← []
for idx in N_modalities do
    docking_output.add(
            DockingLayer[idx](modalities[idx]))
end for
sum_probabilities ← sum(
            probabilities × availability)
probabilities ← probabilities ÷ sum_probabilities
features_idxs ← multinomial(probabilities,
            embracement_size)
output ← sum(docking_output ×
            features_idxs.one_hot(N_modalities))
return output
```

features that then are compacted for output. In our proposal, the probabilities are set in the same value for every modality on each embrace layer, 0.333. However, these probabilities can be configured according to the quality of data and presence of each modality to give more relevance to the more understandable modalities.

The fusion process is shown in Algorithm 2, which corresponds to the process shown in Figure 1(c). The algorithm receives all modalities data: a set of predictions from face images processed at the time of capture, the audio transcribed to text, and processed as MFCC features. First, all models used are instantiated, and also the extra linear layer is added at the end to get the final output (Alg. 2 line 1-3). Then, the availability vector is created; for this, every modality is assessed. That vector is initiated with all 1's; if input data is NULL, the respective value in the vector changes to 0 (Alg. 2 lines 4-10). Afterwards, the probabilities required are set (Alg. 2 line 11); this part can be modified externally. The three fusion models (first EmbraceNet, weighted sum, and concatenation) are used in the first instance (Alg. 2 line 12-15). Finally, the second EmbraceNet merge the outputs of the other methods (Alg. 2 line 18); in principle, this second EmbraceNet use balanced probabilities, but it also can be set by users preferences. An availability vector is not needed for the second EmbraceNet because its inputs will always be available.

Besides, for adding modalities, the EmbraceNet+ method could be easily modified by just adding corresponding docking layers and retraining the whole fusion model with old and added modalities. While the embrace layer and the additional methods (weighted sum and concatenation) will need minor modifications in the instantiation. Regarding the fusion process (Algorithm 2), the availability vector will increase by the number of added modalities (Alg. 2 line 4), and also a corresponding availability supervisor mechanism will be added. The number of probabilities and individual predictions also increase.

### D. MODELING ROBOT BEHAVIOUR

The application of computer systems based on machine learning in robotics promotes the development of technology and a closer relationship with the human being in everyday life. The contribution of these new approaches helps in the development of a more suitable environment for HRI. The detection of emotions through the proposed pre-trained intelligent systems can interfere in the appointed activities to the robot and thus generate greater naturalness in moments of face-to-face or non-verbal communication. The consolidation of this technology in humanoid robots could also help the locomotion process in social environments to directly affect

**Algorithm 2** Fusion Process with EmbraceNet+

```
Require: Facial predictions, transcription, and Mel spectrograms
Instantiate the individual models: $Audio\_model$, $Text\_model$
Instantiate the Fusion methods $EmbraceNet\_1$, and $EmbraceNet\_2$
Adding an extra linear layer of 4 after the $EmbraceNet\_2$
$aval\_1 \leftarrow [1, 1, 1]$
if $facial\_predictions$ == NULL then
    $aval\_1[0] \leftarrow 0$
end if
if transcription == NULL then
    $aval\_1[1] \leftarrow 0$ $text\_prediction \leftarrow Text\_model(transcription)$
end if
if audio == NULL then
    $aval\_1[2] \leftarrow 0$ $audio\_prediction \leftarrow Audio\_model(mean(melspectograms))$
end if
    // The following part corresponds to EmbraceNet+
$prob\_1 \leftarrow SetProbabilities()$
$modalities \leftarrow [mean(face\_predictions), text\_prediction, audio\_prediction]$
$method\_1 \leftarrow EmbraceNet\_1(modalities, aval\_1, prob\_1)$
$method\_2 \leftarrow WeightedSum(modalities, aval\_1, prob\_1)$
$method\_3 \leftarrow Concatenation(modalities, aval\_1)$
$prob\_2 \leftarrow SetProbabilities()$
$methods\_outputs \leftarrow [method\_1, method\_2, method\_3]$
return $EmbraceNet\_2(methods\_outputs)$
```

the speed according to the fear reactions of individuals around the robots. The recognized emotion of a human interacting or not with the robot can support efficient path planning, better social interaction and social navigation, and better performance of the tasks carried out by robots that are at the service of users, such as in health-care homes, hospitals, schools, museums.

## V. EXPERIMENTAL EVALUATION

This section shows the performance of the proposed system compared with two state-of-the-art methods proposed by Poria *et al.* [36] and by Mittal *et al.* [34]. These methods report results with the same dataset we use, making them suitable for comparison with the proposed system, although they are not focused on or applied to social robotics. First, the dataset and its pre-processing are explained; then, the training process of the individual models and the fusion methods are described; afterwards, the applied experiments to validate and evaluate the models and the obtained results are presented. All the aforementioned procedures were developed in a Google Colab environment.

### A. DATASET PRE-PROCESSING

After literature review, the IEMOCAP [65] dataset was chosen to train and test the individual and fusion methods. This dataset is one of the main benchmarks for multimodal emotion recognition that is available with no trouble. IEMOCAP dataset has annotated several data types, such as motion capture face information, speech, videos, head movement and head angle information, dialogue transcriptions, and the word-level, syllable-level and phoneme level alignment. For carrying out the whole pre-processing task (as well as the training and validation procedures), every frame was cut to contain only the annotated person and discard noise data.

Regarding the text modality in IEMOCAP, it was obtained by translating speeches into text, that was performed by human professionals. Although this is not ideal for robotics scenarios, in which such translation is automatically performed by an specialized software application, we used these available data, since IEMOCAP is highly used to train and test emotion recognition models, from which we can compare our approach.

For detection of faces, the RetinaFace model [66] was used because of its accuracy and ease of use. This model is a robust single-stage face detector that performs pixel-based face localization by taking advantage of joint extra-supervised and self-supervised multi-task learning and using a lightweight backbone (ResNet or MobileNet).

When cutting the videos in the annotated sub-videos, several segments were lost because the tools used did not work correctly, causing that the face modality has a much fewer number of samples for training and testing. This also cause that not all samples have the three modalities, but it is certain that all have at least one. However, this fact allows simulating data loss for modality fusion, which helps in the system robustness for its application with data from uncontrolled environments. In addition, the length of videos varies, thus for standardizing the number of frames, we select eight frames per video in a distributed manner.

The available IEMOCAP data (7532 samples) have annotated ten emotion categories (happiness (595), neutral (1708), sadness (1084), anger (1103), excitement (1041), frustration (1849), surprise (107), disgust (2), fear (40), and other (3)), but the number of samples is unbalanced. IEMOCAP's authors suggest altering the category labels to balance them by considering surprise, disgust, fear as the category other, and taking happiness and excitement as the same category due to their similarity in valence, activation, and dominance values [65]. We consider just happiness, neutral, sadness, and anger for our evaluation, discarding frustration as other works to compare. The number of samples per category obeys the following distribution: happiness 1636, neutral 1708, sadness 1084, and anger 1103.

### B. TRAINING OF MODELS

IEMOCAP has its data distributed into five sessions, such that, data from the first four sessions were used for models training, remaining the fifth session just for the test. For facial modality there are 19384 images, obtained after pre-processing (2423 samples × 8 frames each), from which the $\sim$ 78.79% were for training, and the rest ($\sim$ 21.21%) for the test. For audio modality, there are 5531 samples, where the $\sim$ 77.56% were used for training and the $\sim$ 22.44% for the test.

The DialogXL text model, used to analyze text, is pre-trained on the IEMOCAP training dataset, just like the other models in this work and uses six emotions (happiness, neutral, sadness, anger, excitement, and fear). During the original training the AdamW optimizer was used and the tunable hyperparameters were the learning rate, the number of heads for the four types of attentions used in the dialog-aware self-attention component, the max length of memory, and the dropout rate. However, since we decide to work with only four emotions, it was necessary to adapt the text model to this. Thus, we merge happiness with excitement to balance the number of samples in the dataset and the fear-related samples and predictions were removed for the training and for further fusion. Thus, the text modality has 5188 samples, from which $\sim$ 79.76% were used for training and the $\sim$ 20.24% were used for test.

Face and audio models were trained from scratch with the four emotions considered. For training, both models use cross-entropy loss function and the Adam optimizer with a base learning rate set at 0.001, but for the audio model, a scheduler was also used to reduce the learning rate up to 0.000001. The face model presented overfitting, exposed by the increase of validation loss values; thus, the training was stopped in epoch 22 to avoid propagating the errors. The audio model was trained in 60 epochs. The big advantage in this case was the low level of processing when using a neural network with only 3 convolution layers and the data was converted from audio to images with pre-processing.

Finally, the fusion method EmbraceNet+ was trained also with the Adam optimizer and using the cross-entropy loss function. This component takes every available sample from three modalities and the mismatched samples were taken as incomplete, by putting the missing modality as noisy or just absent. Following the number of emotions and the parameters required, the inner EmbraceNet that outputs the final prediction has been configured as [16, 4, 12] for first modified EmbraceNet, weighted sum, and concatenation, respectively, and a linear layer with four neurons. The fusion trained takes 20 epochs. Furthermore, every model (face, audio, text, and EmbraceNet+) were developed, trained, and validated using the PyTorch framework.

### C. EXPERIMENTS AND RESULTS

The three individual models and the entire composite system are evaluated separately. Every evaluation is made with common classification metrics: accuracy and F1 score.

The accuracy of individual models is shown in Table 3. For face modality the accuracy obtained is 44.0%, 58.3% in audio modality, and for text modality is 83.5% (after reducing to four categories). These results show a deficiency in facial mode due to the smaller number of samples and a bit of overfitting with the training samples. Nevertheless, the model used is robust enough to be considered a contributor to our multimodal system. The poor face modality results might be also related to the data nature (videos) and the processing way (selected frames) because some noise data is introduced without a filter or discriminator. Likewise, the improved results for text modality could be influenced by the quality of the data. As mentioned before, the text data in IEMOCAP have been heavily curated and processed by professionals, so these data have less noise than video and audio modalities, which we had to pre-process ourselves. This fact naturally leads to better results in the text modality with respect to the other two.

**TABLE 3. Results of each modality in metrics F1 Score and accuracy (Acc).**

| Emotion Category | Face Modality | | Audio Modality | | Text Modality | |
|---|---|---|---|---|---|---|
| | F1 | Acc (%) | F1 | Acc (%) | F1 | Acc (%) |
| Happiness | 0.70 | 66.8 | 0.44 | 32.1 | 0.85 | 84.9 |
| Neutral | 0.23 | 16.2 | 0.55 | 66.1 | 0.78 | 82.4 |
| Sadness | 0.31 | 21.7 | 0.62 | 69.8 | 0.86 | 81.7 |
| Anger | 0.32 | 71.4 | 0.64 | 65.3 | 0.89 | 84.5 |
| Average | 0.39 | 44.0 | 0.56 | 58.3 | 0.84 | 83.5 |

Hence, according to these experiments, due to the higher results of the text model, this modality could be sufficient on its own for practical applications, as it is also demonstrated by other authors [53]; however, the integration of other modalities is still necessary to provide flexibility and robustness [51], [52]. While results may vary depending on the fusion method used, using other modalities to complement the text input allows the model to deal with ambiguities or samples where there is not enough information given in the text and a context is necessary.

Regarding the multimodal approach, we carried out four experiments related to the absence of one of the modalities: (i) evaluation of all modalities, i.e., face, audio, and text **(F+A+T)**; (ii) evaluation of just face and text modalities **(F+T)**; (iii) evaluation of just audio and text modalities **(A+T)**; and (iv) evaluation of just face and audio modalities **(F+A)**. In all cases, the EmbraceNet+ fusion method is used, which in turn automatically detect the absence of modalities and adjust the probabilities, as shown in Algorithm 1 and Algorithm 2. Figure 6 shows the results of each test in a respective confusion matrix. Except in the F+A evaluation, all tests show a good performance for classifying emotions. Therefore, the presence of text modality has a great impact on the performance of the system. Due to the text is a transcription of the audio data, the text modality only could be with troubles if the audio data also have problems in the recording. Having both audio and text modalities missing or with problems is the worst-case scenario for our proposed system.

To show the results of the proposed approach, a set of comparisons with the results reported by Poria *et al.* [36]

**TABLE 4.** Comparisons between state-of-the-art methods in IEMOCAP dataset [65].

| Method | Poria et al. [36] | | | | Mittal et al. [34] | | Our Method | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Evaluation | F+A+T | F+T | A+T | F+A | F+A+T | | F+A+T | | F+T | | A+T | | F+A | |
| | Accuracy (Acc) (%) | | | | F1 | Acc (%) | F1 | Acc (%) | F1 | Acc (%) | F1 | Acc (%) | F1 | Acc (%) |
| Happiness | 72.2 | 65.2 | 69.2 | 69.4 | 0.86 | 81.6 | 0.82 | 80.1 | 0.85 | 85.6 | 0.82 | 79.4 | 0.54 | 41.9 |
| Neutral | 80.4 | 69.3 | 77.5 | 77.6 | 0.75 | 74.4 | 0.74 | 79.2 | 0.74 | 76.5 | 0.74 | 77.9 | 0.57 | 69.0 |
| Sadness | 75.6 | 63.3 | 74.9 | 74.2 | 0.82 | 88.1 | 0.78 | 76.3 | 0.80 | 73.5 | 0.78 | 78.0 | 0.61 | 64.1 |
| Anger | 79.2 | 62.5 | 74.8 | 71.9 | 0.86 | 86.8 | 0.81 | 74.7 | 0.76 | 77.6 | 0.82 | 76.5 | 0.64 | 64.7 |
| Average | 76.9 | 65.1 | 74.1 | 73.3 | 0.82 | 82.7 | 0.79 | 77.6 | 0.79 | 78.3 | 0.79 | 78.0 | 0.59 | 59.9 |

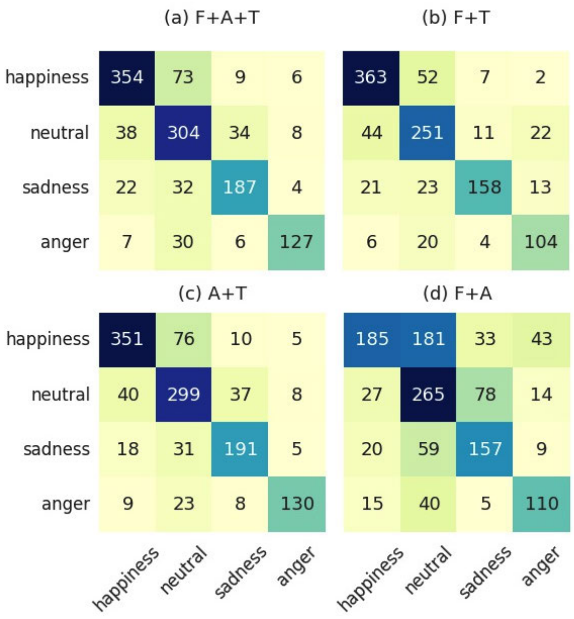


**FIGURE 6.** Confusion matrices of four multimodal evaluation. While the squares are darker, greater and better are the results for the respective emotion category.

(accuracy) and Mittal *et al.* [34] (F1 and accuracy) are carried out, as presented in Table 4 and Figure 7. These methods do not detail how the samples of the happiness category were increased; Poria *et al.* [36] report the use of 1630 samples for happiness, which indicates that they should do some data increase, such as merging happiness and excitement or applying some oversampling technique. The samples for the other three categories coincide with the original samples in IEMOCAP. These experimental conditions are similar to us. Figure 7 shows the reported accuracies per emotions and the average of both Poria *et al.* [36] model and ours. Our proposed system surpasses the model of Poria *et al.* in every test, except in the F+A evaluation. Deleting the audio modality (F+T), our proposal achieves a greater performance, which means that face and text modalities are successfully correlated and complemented. On the other hand, by removing text mode, the system works poorly and only gets it right almost 60 per cent of the time. The best scenario (F+T) also outperforms the more favorable or complete one (F+A+T) by a bit because there are samples with audio and face modalities without text, which makes a little fall in overall results. These evaluations may show a little dependency on the text modality, which is understandable as this modality is the most informative in terms of people's feelings and has a complex and well-established method behind it.

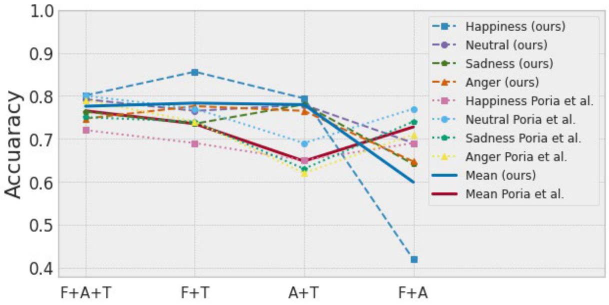


**FIGURE 7.** Graph comparison of reported accuracies (axis y) by category and average over four evaluations (axis x), between our results and poria *et al.* [36].

The method from Mittal *et al.* (M3ER) [34] remains as the better method in the IEMOCAP dataset for the combination F+A+T, with 0.82 of F1 score and 82.7% of accuracy (considering the happiness category), which means 0.03 and 5.1% over our results in the same metrics (see Table 4). Mittal *et al.* [34] do not detail the number of samples used in their study. However, this comparison is just referential because it is unclear how many samples of each category this model use. In particular, this study does not report how the happiness samples were increased to overcome the unbalancing of the original samples in IEMOCAP. We compare only the F+A+T scenario because this work only reports the results using these three modalities combined. Our results are slightly below but competitive regarding neutral, sadness, and anger. For instance, in neutral, we surpass the accuracy in 5.8%, but with the F1 score, [34] is over for 0.1. In the anger category, [34] is better with a significant difference of 0.5 and 12.1% in F1 and accuracy, respectively.

From these results, we deduce that although our method does not achieve a new state-of-the-art, it reaches competitive predictions for some emotion categories. Furthermore, we can argue that the way of processing each modality in our proposal is different from the ones proposed in [34], [36]. In particular, for M3ER model, a loss of data samples for some modality are not reported. Thus, we can assume that that model uses all modalities in each sample; which is a problem in scenarios where eventually some modality data can be missing or with noise and errors. Therefore, this loss of data allows simulating such scenarios and prepare the system for them. We consider that in the implementation for real-world applications, an additional component that discriminates or filters noise and wrong data would be necessary.

## VI. DISCUSSION

The application of our approach, as a proof-of-concept, demonstrates the feasibility and suitability of a flexible robot system able to recognize emotions from interactions with humans, which can be adapted in terms of robot's sensory capacity and accordingly used to adapt the robot's behavior. This experience also gives the opportunity of extracting its current limitations and some lessons learned.

### A. DATASET PRE-PROCESSING

One of the greatest challenges encountered during the development of this work, was obtaining a dataset from which our models could learn. Most datasets were missing at least one of the modalities we tested in our model; for example, VAM [67] and RECOLA [68] consider just audio and visual data, and EmoDB [69] contains only audio data. After we reviewed different datasets, we decided to use the IEMOCAP, a well-known and widely used dataset. There exist a few others, such as CMU-MOSEAS [70] and MELD [71] that we plan to test to better validate our approach in future research. CMU-MOSEAS was released in 2020. This dataset, as well as IEMOCAP, is an emotion repository, but it also includes sentimental data. CMU-MOSEAS covers four languages: English, Spanish, Portuguese, and French. MELD contains multi-party conversations that are more challenging to classify. It uses seven emotions for the annotation, such as: anger, disgust, fear, joy, neutral, sadness, and surprise.

In order to use IEMOCAP, a pre-processing task was required. Since our model uses static images, we had to split each video into individual frames. The audio files then had to be synced up with each individual frame that was used for training; the text file used did not have each utterance uniquely identified which led to confusing results at first. Even after having the data correctly formatted, we still had the problem of imbalanced data, making our networks struggle to classify some emotions.

In our experience, this is a common issue in multi-label emotion classification (i.e., multimodal, multisource), when selecting too many emotions, this lead us to the decision of removing the underrepresented emotions entirely from the dataset. For future studies that require different modalities of data, such as audio and visual data, it is a must the creation of new datasets which takes into consideration the sync among different modalities from the start. We expect that researchers that have these possibilities, can share their datasets for the community interested.

Another issue is regarding the transcription of speech to text in IEMOCAP, as it was done by professionals. Although this professional translation does not represent the robotic scenario, in which it is done by specialized speech-to-text software, the IEMOCAP is a well known and highly used dataset, that made it possible to perform the comparison of our approach with state of the art approaches. Nevertheless, it is also a must to generate datasets with data gathered from the robotic perspective. For example, with the use of our system in a robot, it is possible to use speech-to-text machine translation techniques, for a fast automatic production of new data that could be used for comparison in new experiments. Another way to solve this issue would be the use of off-the-shelf transcription services, such as the IBM Watson Speech-To-Text service.

In summary, to have a better support in the developing of multimodal emotion recognition systems for social robots, more appropriate datasets are need, that consider multiple co-related modalities and generated from a robotic perspective.

### B. ADAPTABILITY AND FLEXIBILITY

As stated in this work, the number and types of modalities are essential in multimodal emotion detection. In a practical social robotics application, different robots may have different sensors or some of them might malfunction during regular use. Therefore, an adaptive approach is highly desired in these scenarios. The proposed architecture represents a step towards this direction, since it adapts to these constraints. As results show, our current fusion architecture allows using as many data sources as available and adapting if any of them is still missing. In the future, we hope to improve the automatic detection of available sources and determination of the importance of each one. For this, more real data samples (with all data samples, missing data, different levels of quality data, etc.) would be needed to pre-train a more robust model.

### C. APPLICABILITY OF THIS ALGORITHM

The flexibility of the approach allows being applied in different contexts where emotions are part of them. As we described before, in social robotics, emotion recognition is commonly used to improve the making decision of robots, modifying it behaviour and thus, the HRI.

The wide study of computer vision, which is a field that deals with how computers can gain high-level understanding from images or videos, has allowed the development of different investigations where the analysis of emotions is taken into account. Research ranging from the analysis of emotions in pedestrians [72], patients [73], [74], car drivers [75], people in restaurants [76], to TV show reactions [77], are examples in which this approach can be also applied.

Figure 8 depicts the pipeline of our current system implementation that can be generalized for any application. Our current version considers images, audio, and text as input data, that can be captured by any appropriate sensor, such as cameras, microphones, keyboards, to process people speaking among them or people interacting with other devices (e.g., screen players, TV, PC), as well as complemented with

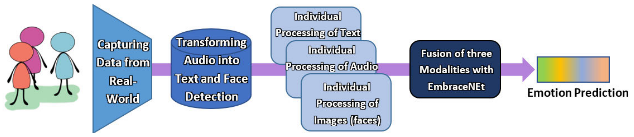

Capturing Data from Real-World
Transforming Audio into Text and Face Detection
Individual Processing of Text
Individual Processing of Audio
Individual Processing of Images (faces)
Fusion of three Modalities with EmbraceNEt
Emotion Prediction

**FIGURE 8.** Pipeline generalized of the proposed system.

other sources or modalities, supported on the adaptability of the EmbraceNet+ fusion model.

### D. EXTENSIBILITY

While the approach presented here is meant for emotion detection in a single person, it could be extended with emotion detection for group of people, to let robots adapt their behaviours accordingly; for example, to decide approaching to a group of people or not, to adapt their navigation to be more socially accepted, by do not traversing among a group of people. This could work also as a base for a group cohesion detection system. Group cohesion can be defined as the measure of bonding between the members of the group [78].

Another use for emotion detection in multiple subjects could be in real time systems (RTS) like security systems, where it can be used to identify if a subject is acting radically differently from the rest of the people in the group.

## VII. CONCLUSION

Social robotics is an active research area in Artificial Intelligent concerned with HRI in a socially acceptable fashion. Robots have to be able of processing the intention from human being and be empowered to react to them. The use of computer vision contributes robot activities in real social environments. Perception and emotion recognition are very important tools for social robotics allowing to improve the making decision process of the robot. However, robots can have a high variety of sensory capacities and might capture the word through different sources and with different levels of quality, which make emotion recognition approaches ineffective and with low accuracy. In this sense, this work proposes a system to analyze the emotions from different modalities and sources, such as facial expressions, text, and audio, integrated by an adaptive fusion method able to manage different modalities and different data quality, in order to better understand the social and behavioral aspects of a human being in a given environment. The proposed system is composed of modules for: the data capturing from real-world; pre-processing strategy of the data inputs that need it, such as transforming audio into text, detecting faces in images; individual processing of each modality (e.g., text, audio, images/faces); and the fusion method. Individual processing of modalities are performed by well-established methods specialized for each modality, that can be easily integrated into the system. To aggregate the individual results, we use the EmbraceNet+ fusion method to merge these modalities and to generate the final prediction, taking into account only present modalities, different levels of data quality, and missing data.

The fusion method is evaluated in four tests combining different modalities: Face-Text (F+T), Face-Audio (F+A), Audio-Text (A+T), and Face-Audio-Text (F+A+T). Results show a good performance surpassing the model of Poria *et al.* [36] in each test except in the F+A evaluation and resting very close to Mittal *et al.* [79] in the test of F+A+T for classifying emotions. Our results reach competitive predictions compared with state of the art approaches for some emotion categories, but with the advantage of being flexible and adaptive. Even though results reveal that the text model could be enough on its own for emotion recognition in practical applications, the integration of other modalities can provide flexibility and robustness. Indeed, using other modalities to complement the text input allows the model to deal with ambiguities or samples when the text does not provide enough information or when a context is required.

We also present in this work an extensive review of state-of-the-art on fusion methods for multimodal emotion recognition approaches. We show that existing approaches are not appropriate for social robots, since they are unable to be adapted according robots' sensory capacity and quality of data. Thus, our approach represent an important contribution in this area.

We are currently working on optimization and tuning the learned models and the pre-processing of inputs for the synchronous emotion detection. The new tests include performance measures of robot responses in real time and real scenarios, with more variety of modalities and considering time-varying data, in which emotions change in time. We also plan to make a formal analysis of time complexity and experiments to measure the total execution time in such real time scenarios; as well as more experiments with other available datasets that consider several modalities, such as CMU-MOSEAS [70] and MELD [71].

## REFERENCES

[1] N. Webb, A. Ruiz-Garcia, M. Elshaw, and V. Palade, "Emotion recognition from face images in an unconstrained environment for usage on social robots," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–8.

[2] J. James, C. I. Watson, and B. MacDonald, "Artificial empathy in social robots: An analysis of emotions in speech," in *Proc. 27th IEEE Int. Symp. Robot Human Interact. Commun. (RO-MAN)*, Aug. 2018, pp. 632–637.

[3] S. Saunderson and G. Nejat, "How robots influence humans: A survey of nonverbal communication in social human–robot interaction," *Int. J. Social Robot.*, vol. 11, no. 4, pp. 575–608, Aug. 2019.

[4] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka, "Social robots for education: A review," *Sci. Robot.*, vol. 3, no. 21, pp. 1–9, Aug. 2018.

[5] S. N. Mohammed and A. K. A. Hassan, "A survey on emotion recognition for human robot interaction," *J. Comput. Inf. Technol.*, vol. 28, no. 2, pp. 125–146, 2020.

[6] B. Kratzwald, S. Ilić, M. Kraus, S. Feuerriegel, and H. Prendinger, "Deep learning for affective computing: Text-based emotion recognition in decision support," *Decis. Support Syst.*, vol. 115, pp. 24–35, Nov. 2018.

[7] L. Chen, Z. Liu, M. Wu, K. Hirota, and W. Pedrycz, "Multimodal emotion recognition and intention understanding in human-robot interaction," in *Developments in Advanced Control and Intelligent Automation for Complex Systems*. Cham, Switzerland: Springer, Mar. 2021, pp. 255–288, doi: 10.1007/978-3-030-62147-6_10.

[8] A. Kumar, K. Srinivasan, W.-H. Cheng, and A. Y. Zomaya, "Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data," *Inf. Process. Manage.*, vol. 57, no. 1, Jan. 2020, Art. no. 102141.

[9] S. Garg, N. Sünderhauf, F. Dayoub, D. Morrison, A. Cosgun, G. Carneiro, Q. Wu, T.-J. Chin, I. Reid, S. Gould, P. Corke, and M. Milford, "Semantics for robotic mapping, perception and interaction: A survey," *Found. Trends Robot.*, vol. 8, nos. 1–2, pp. 1–224, 2020.

[10] M. Spezialetti, G. Placidi, and S. Rossi, "Emotion recognition for human-robot interaction: Recent advances and future perspectives," *Frontiers Robot. AI*, vol. 7, p. 145, Dec. 2020.

[11] G. Recio, W. Sommer, and A. Schacht, "Electrophysiological correlates of perceiving and evaluating static and dynamic facial emotional expressions," *Brain Res.*, vol. 1376, pp. 66–75, Feb. 2011.

[12] T. B. Sheridan, "Human–robot interaction: Status and challenges," *Human Factors*, vol. 58, no. 4, pp. 525–532, Apr. 2016.

[13] B. K. Chakraborty, D. Sarma, M. K. Bhuyan, and K. F. MacDorman, "Review of constraints on vision-based gesture recognition for human–computer interaction," *IET Comput. Vis.*, vol. 12, no. 1, pp. 3–15, Feb. 2018.

[14] C. Tsiourti, A. Weiss, K. Wac, and M. Vincze, "Multimodal integration of emotional signals from voice, body, and context: Effects of (in) congruence on emotion recognition and attitudes towards robots," *Int. J. Social Robot.*, vol. 11, no. 4, pp. 555–573, 2019.

[15] N. J. Shoumy, L.-M. Ang, K. P. Seng, D. M. M. Rahaman, and T. Zia, "Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals," *J. Netw. Comput. Appl.*, vol. 149, Jan. 2020, Art. no. 102447.

[16] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image Vis. Comput.*, vol. 65, no. 1, pp. 3–14, Sep. 2017.

[17] C. Yu and A. Tapus, "Interactive robot learning for multimodal emotion recognition," in *Social Robotics*. Cham, Switzerland: Springer, Nov. 2019, pp. 633–642, doi: 10.1007/978-3-030-35888-4_59.

[18] S. Adiga, D. Vaishnavi, S. Saxena, and S. Tripathi, "Multimodal emotion recognition for human robot interaction," in *Proc. 7th Int. Conf. Soft Comput. Mach. Intell. (ISCMI)*, Nov. 2020, pp. 197–203.

[19] J. Heredia, Y. Cardinale, I. Dongo, and J. Díaz-Amado, "A multi-modal visual emotion recognition method to instantiate an ontology," in *Proc. 16th Int. Conf. Softw. Technol.*, 2021, pp. 453–464.

[20] B. Ko, "A brief review of facial emotion recognition based on visual information," *Sensors*, vol. 18, no. 2, p. 401, Jan. 2018.

[21] N. Mehendale, "Facial emotion recognition using convolutional neural networks (FERC)," *Social Netw. Appl. Sci.*, vol. 2, no. 3, pp. 1–8, Mar. 2020.

[22] M. Omkar Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2015, pp. 41.1–41.12.

[23] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," *Sensors*, vol. 21, no. 9, p. 3046, Apr. 2021.

[24] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1724–1734.

[25] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," 2014, *arXiv:1409.1259*.

[26] T. Wolf *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," 2019, *arXiv:1910.03771*.

[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[28] B. Krause, L. Lu, I. Murray, and S. Renals, "Multiplicative LSTM for sequence modelling," 2016, *arXiv:1609.07959*.

[29] T. Wolf *et al.*, "Transformers: State-of-the-art natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations*. Stroudsburg, PA, USA: Association for Computational Linguistics, Oct. 2020, pp. 38–45.

[30] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech emotion recognition using Fourier parameters," *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 69–75, Jan. 2015.

[31] L. Zheng, Q. Li, H. Ban, and S. Liu, "Speech emotion recognition based on convolution neural network combined with random forest," in *Proc. Chin. Control Decis. Conf. (CCDC)*, Jun. 2018, pp. 4143–4147.

[32] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. Interspeech*, Sep. 2014, pp. 223–227.

[33] S. Tripathi, S. Tripathi, and H. Beigi, "Multi-modal emotion recognition on IEMOCAP dataset using deep learning," 2018, *arXiv:1804.05788*.

[34] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3ER: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in *Proc. AAAI*, vol. 34, no. 2, Apr. 2020, pp. 1359–1367.

[35] J. D. S. Ortega, M. Senoussaoui, E. Granger, M. Pedersoli, P. Cardinal, and A. L. Koerich, "Multimodal fusion with deep neural networks for audio-video emotion recognition," 2019, *arXiv:1907.03196*.

[36] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Dec. 2016, pp. 439–448.

[37] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski, and R. C. Ferrari, "EmoNets: Multimodal deep learning approaches for emotion recognition in video," *J. Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.

[38] P. Tzirakis, J. Chen, S. Zafeiriou, and B. Schuller, "End-to-end multimodal affect recognition in real-world environments," *Inf. Fusion*, vol. 68, pp. 46–53, Apr. 2021.

[39] M. S. Akhtar, D. Chauhan, D. Ghosal, S. Poria, A. Ekbal, and P. Bhattacharyya, "Multi-task learning for multi-modal emotion recognition and sentiment analysis," in *Proc. Conf. North*, Jun. 2019, pp. 370–379.

[40] B. Sun, Q. Xu, J. He, L. Yu, L. Li, and Q. Wei, "Audio-video based multimodal emotion recognition using SVMs and deep learning," in *Pattern Recognition*. Singapore: Springer, Oct. 2016, pp. 621–631.

[41] Y.-T. Lan, W. Liu, and B.-L. Lu, "Multimodal emotion recognition using deep generalized canonical correlation analysis with an attention mechanism," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–6.

[42] W. Li, M. Chu, and J. Qiao, "Design of a hierarchy modular neural network and its application in multimodal emotion recognition," *Soft Comput.*, vol. 23, no. 22, pp. 11817–11828, Nov. 2019.

[43] S.-F. Zhang, J.-H. Zhai, B.-J. Xie, Y. Zhan, and X. Wang, "Multimodal representation learning: Advances, trends and challenges," in *Proc. Int. Conf. Mach. Learn. Cybern. (ICMLC)*, Jul. 2019, pp. 1–6.

[44] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. EMNLP*, Copenhagen, Denmark, Sep. 2017, pp. 1103–1114.

[45] J.-H. Choi and J.-S. Lee, "EmbraceNet: A robust deep learning architecture for multimodal classification," *Inf. Fusion*, vol. 51, pp. 259–270, Nov. 2019.

[46] V. Vielzeuf, A. Lechervy, S. Pateux, and F. Jurie, "CentralNet: A multilayer approach for multimodal fusion," in *Computer Vision—ECCV Workshops*. Cham, Switzerland: Springer, Sep. 2018, pp. 575–589.

[47] N. Audebert, C. Herold, K. Slimani, and C. Vidal, "Multimodal deep networks for text and image-based document classification," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Würzburg, Germany: Springer, 2019, pp. 427–443.

[48] J. Venugopalan, L. Tong, H. R. Hassanzadeh, and M. D. Wang, "Multimodal deep learning models for early detection of Alzheimer's disease stage," *Sci. Rep.*, vol. 11, no. 1, pp. 1–13, Dec. 2021.

[49] Y. R. Pandeya, B. Bhattarai, and J. Lee, "Deep-Learning-Based multimodal emotion classification for music videos," *Sensors*, vol. 21, no. 14, p. 4927, Jul. 2021.

[50] Y. R. Pandeya and J. Lee, "Deep learning-based late fusion of multimodal information for emotion classification of music video," *Multimedia Tools Appl.*, vol. 80, no. 2, pp. 2887–2905, Jan. 2021.

[51] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," 2019, *arXiv:1906.00295*.

[52] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," 2018, *arXiv:1811.09362*.

[53] W. Rahman, M. K. Hasan, S. Lee, A. B. Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," *ACL Anthol.*, vol. 2020, pp. 2359–2369, Jul. 2020.

[54] K. Liu, Y. Li, N. Xu, and P. Natarajan, "Learn to combine modalities in multimodal deep learning," 2018, *arXiv:1805.11730*.

[55] R. Stock-Homburg, ''Survey of emotions in human–robot interactions: Perspectives from robotic psychology on 20 years of research,'' *Int. J. Social Robot.*, pp. 1–23, Jun. 2021.

[56] R. Savery and G. Weinberg, ''A survey of robotics and emotion: Classifications and models of emotional interaction,'' in *Proc. 29th IEEE Int. Conf. Robot Human Interact. Commun. (RO-MAN)*, Aug. 2020, pp. 986–993.

[57] N. Rawal and R. M. Stock-Homburg, ''Facial emotion expressions in human-robot interaction: A survey,'' 2021, *arXiv:2103.07169*.

[58] J. Ginés, F. Martín, D. Vargas, F. J. Rodríguez, and V. Matellán, ''Social navigation in a cognitive architecture using dynamic proxemic zones,'' *Sensors*, vol. 19, no. 23, p. 5189, Nov. 2019.

[59] A. Kashii, K. Takashio, and H. Tokuda, ''Ex-amp robot: Expressive robotic avatar with multimodal emotion detection to enhance communication of users with motor disabilities,'' in *Proc. 26th IEEE Int. Symp. Robot Hum. Interact. Commun. (RO-MAN)*, Aug. 2017, pp. 864–870.

[60] W. Graterol, J. Diaz-Amado, Y. Cardinale, I. Dongo, E. Lopes-Silva, and C. Santos-Libarino, ''Emotion detection for social robots based on nlp transformers and an emotion ontology,'' *Sensors*, vol. 21, no. 4, pp. 1–19, 2021.

[61] A. U. Nasib, H. Kabir, R. Ahmed, and J. Uddin, ''A real time speech to text conversion technique for Bengali language,'' in *Proc. Int. Conf. Comput., Commun., Chem., Mater. Electron. Eng. (IC4ME)*, Feb. 2018, pp. 1–4.

[62] K. Venkataramanan and H. R. Rajamohan, ''Emotion recognition from speech,'' 2019, *arXiv:1912.10458*.

[63] W. Shen, J. Chen, X. Quan, and Z. Xie, ''DialogXL: All-in-one XLNet for multi-party conversation emotion recognition,'' 2020, *arXiv:2012.08695*.

[64] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, ''XLNet: Generalized autoregressive pretraining for language understanding,'' in *Proc. NeurIPS*, vol. 32, 2019, pp. 1–11.

[65] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, ''IEMOCAP: Interactive emotional dyadic motion capture database,'' *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

[66] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, ''RetinaFace: Single-stage dense face localisation in the wild,'' 2019, *arXiv:1905.00641*.

[67] M. Grimm, K. Kroschel, and S. Narayanan, ''The vera am Mittag German audio-visual emotional speech database,'' in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jun. 2008, pp. 865–868.

[68] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, ''Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions,'' in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–8.

[69] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, ''A database of German emotional speech,'' in *Proc. INTERSPEECH*, Lisbon, Portugal, vol. 5, Sep. 2005, pp. 1517–1520.

[70] A. Bagher Zadeh, Y. Cao, S. Hessner, P. P. Liang, S. Poria, and L.-P. Morency, ''CMU-MOSEAS: A multimodal language dataset for spanish, portuguese, German and French,'' in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, p. 1801.

[71] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, ''MELD: A multimodal multi-party dataset for emotion recognition in conversations,'' 2018, *arXiv:1810.02508*.

[72] S. Cœugnet, B. Cahour, and S. Kraiem, ''Risk-taking, emotions and socio-cognitive dynamics of pedestrian street-crossing decision-making in the city,'' *Transp. Res. F, Traffic Psychol. Behaviour*, vol. 65, pp. 141–157, Aug. 2019.

[73] A. A. Zainuddin, S. Superamaniam, A. C. Andrew, R. Muraleedharan, J. Rakshys, J. Miriam, M. A. S. M. Bostomi, A. M. A. Rais, Z. Khalidin, A. F. Mansor, and M. S. M. Taufik, ''Patient monitoring system using computer vision for emotional recognition and vital signs detection,'' in *Proc. IEEE Student Conf. Res. Develop. (SCOReD)*, Sep. 2020, pp. 22–27.

[74] M. A. R. Ahad, A. D. Antar, and O. Shahid, ''Vision-based action understanding for assistive healthcare: A short review,'' in *Proc. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2019, pp. 1–11.

[75] M. Oehl, W. F. Siebert, T.-K. Tews, R. Höger, and H.-R. Pfister, ''Improving human-machine interaction—A non invasive approach to detect emotions in car drivers,'' in *Human-Computer Interaction. Towards Mobile and Intelligent Interaction Environments*, J. A. Jacko, Ed. Berlin, Germany: Springer, 2011, pp. 577–585.

[76] D. Sivabalaselvamani and B. Soorya, ''Convolution neural network based specialized restaurant rating using facial expression detection,'' in *Proc. Int. Conf. Inventive Comput. Technol. (ICICT)*, Feb. 2020, pp. 739–744.

[77] S. M. Zahiri and J. D. Choi, ''Emotion detection on tv show transcripts with sequence-based convolutional neural networks,'' in *Proc. Workshops 32nd aaai Conf. Artif. Intell.*, 2018, pp. 44–51.

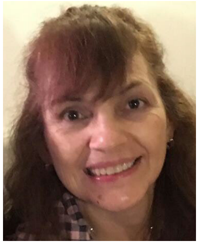
[78] S. Ghosh, A. Dhall, N. Sebe, and T. Gedeon, ''Predicting group cohesiveness in images,'' in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.

[79] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, ''EmotiCon: Context-aware multimodal emotion recognition using Frege's principle,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14234–14243.

**JUANPABLO HEREDIA** received the B.Sc. degree in computer science from Catholic San Pablo University, Arequipa, Peru, in 2020.

He has been participating as a Research Student in the Project Robots for Urban Tourism Centers, Autonomous and Semantic-Based (RUTAS), since April 2020, where he developed his undergraduate thesis and participated in other research. His research interests include machine/deep learning models, computer vision algorithms, graph-based machine/deep learning models, affective computing, and other neuroinformatics topics.

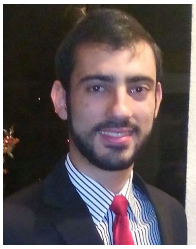
**EDMUNDO LOPES-SILVA** was born in Vitória da Conquista, Bahia (BA), Brazil, in 1998. He is currently pursuing the degree in electrical engineering with the Federal Institute of Bahia, Brazil. He is an Active Member of the Automation and Robotics Innovation and Research Group (GIPAR) and a member of the Chapter of the Institute of Electrical and Electronics Engineers—Robotics and Automation Society (IEEERAS). His research interests include the area of electrical engineering, with emphasis on robotics, acting in robotic perception systems, social robots, machine/deep learning, computer vision, and applications control with artificial intelligence techniques. He has been participating as a Research Student in the Project Robots for Urban Tourism Centers, Autonomous and Semantic-Based (RUTAS), since April 2020.



**YUDITH CARDINALE** received the Graduate degree in computer engineering from Universidad Centro-Occidental Lisandro Alvarado (UCLA), Venezuela, in 1990, and the M.Sc. and Ph.D. degrees in computer science from Universidad Simón Bolívar (USB), Venezuela, in 1993 and 2004, respectively. She has been a Full Professor with the Computer Science Department, USB, since 1996. She is currently an Associate Researcher with Universidad Católica San Pablo, Arequipa, Peru. She has written a range of scientific articles published in international journals, books, and conferences, and has participated as a member of program committees of several international conferences and journals. Her research interests include parallel processing, distributed object processing, operating systems, digital ecosystems, high performance on grid and cloud platforms, collaborative frameworks, and web services composition, including semantic web.

**JOSE DIAZ-AMADO** received the Graduate degree in electronic engineering from University Católica Santa Maria (UCSM), Peru, in 2003, the master's and Ph.D. degrees in electrical and computer engineering from the Federal University of Rio Grande do Norte (UFRN), Brazil, in 2008 and 2013, respectively, and the Postdoctoral degree from the Institute of Mathematical and Computer Sciences, University of Sao Paulo (USP), Brazil, in 2019. He is currently an Associate Researcher with the University of Cátolica San Pablo (UCSP), Peru, and an Adjunct Professor with the Federal Institute of Education, Science and Technology of Bahia (IFBA), Brazil. Tutor of the IFBA Student Chapter the Institute of Electrical and Electronics Engineers—Society of Robotics and Automation (IEEE-RAS). His research interests include the area of electrical engineering, with an emphasis on robotics, acting mainly in the following lines of research, such as autonomous navigation systems for mobile robotics, robotic perception systems, SLAM/location robotic systems, social robots, knowledge representation, deep learning, computer vision, and applications control with artificial intelligence techniques.

**IRVIN DONGO** received the B.Sc. degree in computer science from Catholic San Pablo University, Peru, in 2012, and the M.Sc. and Ph.D. degrees from the University of Pau, France, in 2014 and 2017, respectively. He was a Postdoctoral Fellow with the École Supérieure des Technologies Industrielles Avancées (ESTIA), from 2018 to 2020, France. He is currently an Associate Researcher in computer science with the École Supérieure des Technologies Industrielles Avancées and Catholic San Pablo University. His research interests include normalization and anonymization of web resources, knowledge-bases modeling (semantic web), policies and management of credentials, security model and anonymization technique, machine/deep learning techniques for an analysis and classification of data to discover patters, gesture recognition, and affective computing.



**WILFREDO GRATEROL** is currently pursuing the B.Sc. degree in computer science with Simon Bolivar University, Venezuela. He is also working in his undergraduate dissertation and is an Active Member of the University's Artificial Intelligence Group (GIA) and a Former Member of the Future Robotics (Futbot) Team, where he has participated in developed of different robots and machine learning models for national and international competitions. His research interests include natural language processing (NLP), computer vision, social robotics, machine/deep learning techniques for an analysis and classification of data to discover patterns and deliver recommendations in applications, and real time feedback in social robots.

**ANA AGUILERA** received the B.S. degree (Hons.) in computer science engineering from Lisandro Alvarado Central Western University (UCLA), Barquisimeto, Venezuela, in 1994, the M.S. degree in computer science from Simon Bolivar University, Caracas, Venezuela, in 1998, and the Ph.D. degree in medical informatics from the University of Rennes I, Rennes, France, in 2008. She is currently a Full Professor with the Faculty of Engineering, Escuela de Ingeniería Informática, University of Valparaíso, Valparaíso, Chile. Her research interests include the fuzzy databases, data mining, social networks, and medical informatics. She was accredited in Program for Researcher Promotion of Venezuela, Candidate Level, in 1998. Since 2011, she has been a member of the Program Encouragement for Research and Innovation Researcher (PEII) Level C, Venezuela. She is a member of the Venezuelan Association for the Advancement of Science (AsoVAC) and a member of Venezuelan Computer Society (SVC). She received the Magna Cum Laude Award for B.Sc. degree from UCLA and "Très Honorable" Award in the Ph.D. thesis from Rennes I.