Towards the Evaluation of the Role of Embodiment in Emotions Elicitation

1st Silvia Rossi
Department of Electrical Engineering and Information Technologies
University of Naples Federico II
Napoli, Italy
silvia.rossi@unina.it

2nd Alessandra Rossi
Department of Electrical Engineering and Information Technologies
University of Naples Federico II
Napoli, Italy
alessandra.rossi@unina.it

3rd Sara Sangiovanni
Department of Electrical Engineering and Information Technologies
University of Naples Federico II
Napoli, Italy

***Abstract*—In the human-robot interaction (HRI) literature, there is a focus on the ability of a robotic system to display emotional expressions to make them easily recognizable by the users. However, only few studies addressed how the multimodal behaviours of a robot may cause a specific emotional reaction in a user. In this work, we deploy a late fusion model for emotion recognition using facial and bio-signal inputs, and we use such a model in a user study aiming at evaluating the role of embodied interaction in emotion elicitation. To this extent, we evaluate the ability of a robot to change the elicited emotion with respect to the one elicited by a standardized stimulus (e.g., a video). Initial results show that the robot's behaviour is effective in changing the valence of the perceived emotion, but not the arousal that is lowered.**

***Index Terms*—emotion elicitation, human-robot interaction, non-verbal behaviour, multi-modal fusion**

I. Introduction

Emotions pervade our daily lives and are of extraordinary importance in the context of psychological processes. In human-robot interaction, it is important that the robot recognizes and is able to simulate people's emotions, creating a natural and empathic interaction between people and robots. Indeed, Social Robots (SR) to be effective need to be able to both correctly perceive the emotional reactions of the users, but also to adapt their assistive behaviours to each user [1]. A robot that has the human innate capability to show empathy can provide effective interactions, such as to sustain children's emotional status in vaccination centres [2], and even in paediatric Emergency Departments (ED) [3] as they are designed to increase motivation and sustain engagement while displaying emotional behaviours. Non-verbal communications of a robot through body movements have been investigated in [4] aiming at evaluating emotion elicitation in users during a lecture taken by a NAO robot. While the students were not able to recognize differences in the mood of the robot, the self-reported value seemed to align with it.

In the HRI literature, particular attention is directed on endowing a robotic system with the ability to display emotional expressions in a multimodal way to make them easily recognizable by the users [5], [6]. Several multimedia content is used to elicit emotional responses in people by observing their body, eyes (e.g., observing the pupil dilation and contraction) and the overall changes in the facial expressions [7], [8]. However, only a few studies addressed how the multimodal behaviours of a robot may cause a specific emotional reaction in a user while interacting with a robot. For example, in [1], a robot's emotion elicitation capabilities were measured by using electrocardiography (ECG). Elicitation mechanisms were achieved by the means of combining the Pepper robot's body movements (classified in terms of valence and arousal) with coordinated music. However, there is no way to discriminate whether is the robot, its movements, or the music to mainly contribute to the emotion elicitation.

While in Human-Computer Interaction (HCI) benchmarks for emotion elicitation exist using standardized stimuli, in HRI such evaluations in a controlled setting do not exist. Among the multimedia data, the most used for emotion elicitation are videos [7]. For example, Soleymani et al. [9] presented a user-independent emotion recognition method with the aim of retrieving affective tags for videos using the electroencephalogram. Existing methodologies use stimuli that the user watch (videos or images) or hear (music and sounds) to elicit an emotion [10], [11]. These methodologies, while useful for benchmarking Emotion Recognition (ER) capabilities, are not sufficient to provide insight into how the behaviour alone of a robot may elicit such emotions, since affect is also directly related to the interaction with the robot [12]. As a consequence, while such standardized stimuli may be used for benchmarking ER systems, how to isolate the role of the robot embodiment from such stimuli is still unclear.

Here, we present a user study where the role of a robot embodiment is evaluated in terms of its contribution to the elicitation of emotional reactions. For this purpose, a benchmark task is proposed relying on videos to be displayed on the robot screen, and the robot's actions are evaluated in terms of their ability to contribute to such emotional reaction. The videos are selected from a standard dataset used for ER applications. A multimodal emotion recognition system is developed relying on facial expressions and bio-signals such

This work has been supported by Italian MUR and by the European Union's CHIST-ERA project COHERENT (PCI2020-120718-2), and Italian PON R&I 2014-2020 - REACT-EU (CUP E65F21002920003).

979-8-3503-2745-8/23/$31.00 ©2023 IEEE

as electrocardiography and Galvanic Skin Response (GSR). Decision-level fusion (late fusions) is adopted to provide more flexibility in the change of individual channels and single classifiers. The advantage of decision fusion over feature fusion is to easily employ an optimal weighting scheme to adjust the degree of the contribution of each modality to the final decision result based on the reliability of individual modalities [13].

The proposed approach consists in first assessing the emotional reaction elicited by a standardized video to be shown on the tablet of a robot not moving. This is to set the initial baseline for the evaluation of the embodiment contribution. Then, to assess the contribution of the performed non-verbal expression in the elicitation process, the same videos are shown as accompanied by body movements (already evaluated as clearly recognizable) showing an emotional behaviour in coherence or not coherence with the emotion elicited by the video [14]. A similar setting was investigated by Fiorini et al. [10] that used a set of 60 images, retrieved from a standardized database, for eliciting emotions in their participants while they interacted with a Pepper robot showing three emotions (i.e. positive, negative, and neutral). Their aim was to evaluate the contribution of the robot's behaviour in the recognition performance of different algorithms to be used for ER from facial expressions. Results showed that the multimodal behaviours of the robot helped in the recognition of the emotion that was elicited by the images. The main emotion is induced by the image and there is no way to effectively evaluate the contribution of the robot's behaviours. For this reason, we propose a more ecological setting where the robot's behaviours are accompanied by videos to elicit emotion. While pictures sometimes can be more effective than videos in eliciting emotions [15], robot behaviours have a duration in time. Since the human perception of affect is affected by visual motion [16], we believe that in HRI the use of videos as a baseline for evaluation is more suitable. Our goal is to understand whether the robot's behaviour itself is able to induce emotion, and we propose to evaluate such ability by measuring it in shifting the emotion away from the one elicited by the videos.

The proposed approach has been tested on 30 participants in a user study showing that the behaviour of the robot is able to modify the elicited emotion, especially in terms of a change in valence. Summarizing, the contributions of this work are:

- Proposing a benchmark task for the evaluation of the impact of a robot's non-verbal behaviour in eliciting emotions during HRI;
- The development of a multi-modal architecture for emotion recognition that relies on facial expressions and bio-signals and is easily configurable for different inputs.

II. Methods

We propose a system to predict people's emotions by analysing different human stimuli. Facial expressions are one of the most significant non-verbal modalities to express emotions and intentions [17]. However, some people are able to mask their facial emotions by adopting a neutral expression and using non-intuitive human body language that may lead to misinterpretation [18]. For this reason, here, we also want to consider the more reliable physiological signals. These signals are harder to be covered or altered by human disguising [19], and they can be continuously collected, reflecting people's emotions and changes in the emotions in their daily activities. Physiological signals are largely used in clinical diagnostics, and in HRI studies to increase the accuracy and robustness of the emotion recognition system [20], [21]. In this perspective, multimodal systems have a key role in improving the performances of emotion recognition with respect to single-modality approaches in HRI [6].

The most used signals are electromyogram (EMG), electroencephalography (EEG), electrocardiography, galvanic skin response (GSR), skin temperature (ST), skin conductivity (SC), respiration (RSP), body expression, and blood oxygen saturation (OXY). However, using too many bio-signals to recognize human emotions is not suitable for practical applications, and it may hinder people during daily life activities [22], [23]. In this work, we decided to use ECG, and GSR signals because they are both good indicators for the recognition of emotions [24]. Moreover, while single-modal models struggle to assess people's affective states [25], we decided to rely on multimodal emotion recognition. In particular, we propose a multi-modal model trained on facial, ECG and GSR signals for recognising people's emotions. Our model is validated on the AMIGOS dataset [26] on 4 classes of emotion, and it first uses three single-modal classifiers, one each for the types of input signal, and then, we applied a late fusion with a greedy algorithm.

A. The Dataset

The AMIGOS dataset, whose acronym is "A dataset for Multimodal research of affects, personality traits and mood in Individuals and GROUPS", because it contains people's affective responses, through neurophysiological signals and their relationship with personality, mood, social context, and duration of stimuli. Unlike other databases, the AMIGOS dataset includes people's aroused affection in short and long videos in two social contexts: (i) when people are alone (individual setting), and (ii) when they are part of an audience (group setting). In fact, the data is collected in two experimental settings: in the first, 40 participants watched 16 short emotional videos (duration $<$ 250 s); in the second, they watched 4 long videos (duration $>$ 14 min), some of them alone and the others in groups. Since it is more likely that a long video elicits diverse emotional states according to the scenes presented, we selected 16 short videos. The videos' data included Electroencephalogram, Electrocardiogram and Galvanic Skin Response data, which were recorded using wearable sensors. The dataset also includes participants' frontal HD video, and both RGB and depth full-body videos. Affective levels of the participants were reported in a self-assessment (excitement, valence, dominance, sympathy, familiarity and seven basic emotions) and in an external annotation (arousal and valence). The five dimensions are measured on a scale from 1 (low) to 9

(high), and the basic emotions (neutrality, disgust, happiness, surprise, anger, fear, and sadness) are binary values. Moreover, videos were annotated using continuous and change from -1 (low valence/arousal) to 1 (high valence/arousal) scales.

B. Train Data Classification

Considering a classification of emotions using Russell's Circumplex Model [27], we identified 4 classes of emotion given by the 4 quadrants of the model:

- LALV (0): Low Arousal - Negative Valence;
- LAHV (1): Low Arousal - Positive Valence;
- HALV (2): High Arousal - Negative Valence;
- HAHV (3): High Arousal - Positive Valence.

We used the dataset AMIGOS with FACE, ECG and GSR for classifying 16 short videos by quadrants of valence and arousal (high and low). We used a k-means classification method to define the four clusters with thresholds for the labels of arousal and valence [28].

C. Face Classifier

We used a CNN, called AlexNet network [29], for the face classification. The network is composed of::

- The first convolutional layer takes input images of dimensions of 224×224×3 and applies 96 kernels of size 11×11×3;
- The second layer takes the output of the previous level as input and filters it with 256 kernels of size 5×5×48;
- The third and fourth layers have 384 kernels of size 3×3×256 each. These layers have no pooling operations either normalization;
- The fifth layer has 256 kernels of size 3×3×192;
- The last three fully connected layers have 4096, 4096 and 1000 nodes respectively;
- The last layer has 4 neurons and uses the Softmax as the activation function.

We extracted one frame for a second from the video in the AMIGOS dataset, and then we used mini-batch training due to the high dimensions of the dataset. Each mini-batch contains about 7000 images, which are divided into a training set (70%) and a test set (30%). The training data is further divided into a training set (70%) and a validation set (30%). Each mini-batch has been trained with the Adam optimizer [30] with a learning rate of 0.001 and 7 epochs.

D. Bio-signal Classifiers

Biosignal classification has been done using a Support Vector Machine (SVM) network architecture for the ECG and GSR biosignals. As a first step, we pre-processed the data by normalising it through its standard deviation and mean. Then, we extracted only the significant characteristics from each signal that we intended to use for classification. For the classification, we divided the dataset AMIGOS into a training set (70%) and test set (30%), and the training data is further divided into a training set (70%) and validation set (30%).

1) Pre-processing ECG: For pre-processing ECG signals and selecting the features, we applied frequency filters to clean the signal from noise (i.e., Butterworth filter) and identify the position of the R peaks within the signal. Then, we selected the relevant characteristics. In particular, we were interested in evaluating the variation in heart rate (HRV), represented by the difference in R-R intervals e by the consequent instantaneous variations of HR. Heart rate variability or R-R sequence variability can be calculated with different types of analyses:

- Time Domain analysis: this analysis uses the standard deviation of the R-R intervals (SDRRI), which represents the distance of each heartbeat and the root mean square of the successive differences between each heartbeat (RMSSD).
- Frequency Domain analysis: it allows us to highlight some significant components, in this case, the number of beats is not considered but the recurrence rate. The following measurements were analysed:
  - **HF** (High Frequency): is a sign of vagal activation, represents a respiratory component that reflects parasympathetic respiration;
  - **LF** (Low Frequency): is associated with the activation of sympathetic and vagal. It is associated with the baroreflex phenomenon, it reflects sympathetic and parasympathetic modulation;
  - **LF/HF**: represents the sympathovagal balance and is given by the ratio of low-frequency power to high-frequency power.

2) Pre-processing GSR: As for the ECG signals, we first applied frequency filters to clean the signal from noise (i.e., Butterworth filter) and identified the position of the SCRs peaks within the signal; then, we extracted significant characteristics. In particular, we considered the following factors:

- **SCR amplitude**: Amplitude difference in GSR level between SCR onset and the SCR peak;
- **SCR rise time**: The time difference between SCR onset and peak;
- **SCR half recovery time**: Time difference between when the GSR level was recovered to 50% of the SCR amplitude and the peak time.

E. Model Evaluation

The evaluation of the individual classifiers is performed completely in offline mode. In particular, we used the metrics for accuracy, precision, recall and F-measure.

Table I show the performance of the individual implemented networks (FACE, ECG and GSR). We can observe that FACE and ECG had similar performances, with FACE with slightly higher accuracy than ECG. Both FACE and ECG has definitely better performances than GSR.

F. Fusion Module

Previous studies on emotion recognition focused on the use of single sensor modality, features and classifiers, which are ineffective to discriminate complex emotion classes. The

TABLE I
EVALUATION OF THE INDIVIDUAL CLASSIFIERS (FACE, ECG AND GSR) USING THE AMIGOS DATASET.

FACE				
	LALV(0)	LAHV(1)	HALV(2)	HAHV(3)
Precision	0.70	0.71	0.76	0.79
Recall	0.78	0.75	0.72	0.72
F1 Score	0.78	0.76	0.77	0.74
Accuracy	0.77			

ECG				
	LALV(0)	LAHV(1)	HALV(2)	HAHV(3)
Precision	0.75	0.75	0.77	0.78
Recall	0.79	0.76	0.74	0.77
F1 Score	0.77	0.76	0.75	0.78
Accuracy	0.76			

GSR				
	LALV(0)	LAHV(1)	HALV(2)	HAHV(3)
Precision	0.65	0.61	0.73	0.66
Recall	0.69	0.63	0.61	0.71
F1 Score	0.67	0.62	0.67	0.68
Accuracy	0.66			

fusion of multiple modalities aims at improving classification accuracy by exploiting the complementarity of different modalities [13].

An interesting advantage of decision fusion over feature fusion is that we can easily employ an optimal weighting scheme to adjust the degree of the contribution of each modality to the final decision result based on the reliability of individual modalities [13]. Based on the latter observation, we used the following weighting scheme to determine the final output. For a given test data X, the decision output of the fusion system is:

$$c^* = argmax_i \left\{ \prod_{m=1}^{M} P_i(X \mid \lambda_m)^{a_m} \right\} \quad (1)$$

where M is the number of modalities, λ_m is the classifier for the m-th modality, and $P_i(X \mid \lambda_m)$ is its output for the i-th class. The weights a_m which satisfy $0 \leq a_m \leq 1$ and $\sum_{m=1}^{M} a_m = 1$ represent the modality's reliability, which determines its degree of contribution to the final decision.

To set the weights, a greedy algorithm has been used. A greedy algorithm is any algorithm that follows the problem-solving heuristic of making the locally optimal choice at each stage. In many problems, a greedy strategy does not produce an optimal solution, but a greedy heuristic can yield locally optimal solutions that approximate a globally optimal solution in a reasonable amount of time. In this case, the optimal weights are estimated by an exhaustive search of the grid space, where each weight is increased from 0 to 1 with a step size of 0.01 and for each i, the weights producing the best $P_i(X \mid \lambda_m)^{a_m}$, are selected.

III. BENCHMARKING EMBODIED INTERACTION FOR AFFECTIVE ELICITATIONS

In this work, we are interested in evaluating the effect of the robot's embodiment and emotional gesture on affective elicitation. To do so, we propose a benchmarking task where the contribution of the robot's behaviours in eliciting emotions is compared to a baseline with standardized stimuli and no robot contribution. In details, we propose to evaluate the robot contribution by evaluating the ability of a robot to reinforce the emotion induced by the stimulus or modify it toward a different emotional class.

TABLE II
MAPPING OF THE ROBOT'S EMOTIONAL BEHAVIOURS (I.E., SADNESS, JOY, FRUSTRATION) AND RELATIVE CHANGE OF COLOUR OF THE LEDS (I.E., YELLOW, ORANGE, RED, WHITE).

Class Emotion	Coherent	Incoherent	Neutral
LALV	Sad, orange	Joy, yellow	No-emotion, white
LAHV	Joy, yellow	Frustrated, red	No-emotion, white
HALV	Frustrated, red	Joy, yellow	No-emotion, white
HAHV	Joy, yellow	Sad, orange	No-emotion, white

A. The Robot

In our study case, we used an Aldebaran Robotics Pepper[1] robot Y20 V18A. Using the Choregraphe and NAOqi software of the robot, we modelled three types of emotional behaviours of the robot (Coherent, Incoherent, and Neutral) by changing the colour of the eyes' LEDs, body movements, and head and body pose. For each class of emotion (i.e., low and high valence and arousal, and a neutral), we used pre-defined Pepper's gestures which were chosen according to the evaluation of valence and arousal of the affective gestures made by the robot in Marpena et al. study [31].

We modelled the Coherent and Incoherent behaviours with non-verbal gestures that belong to a class quadrant of Russell's Circumplex Model. The non-verbal behaviours described are combined with selected emotional videos (see Table II) that are shown on Pepper's tablet during the interaction.

B. The videos

We chose to show videos from DECAF, which is a multimodal dataset for decoding user physiological responses to affective multimedia content [32].

We decided to use short videos to avoid having more than one emotion. Table III shows the videos selected for each class

[1]Aldebaran Robotics https://www.aldebaran.com/en/pepper

TABLE III
VIDEO SELECTED FOR EACH CLASS.

Class	Video	Duration	Scene Description
LALV	Bambi	90.1s	Bambi's mother gets killed
	UP	89.1s	Old Carl loses his wife
	Life is Beautiful	112.1s	Guido shot to death by a Nazi solder
LAHV	UP	67.1s	Carl, a shy and quiet boy meets the energetic Elle
	August Rush	90.1s	A son meets his lost mother
	Wall-e	90.2s	Wall-E and Eve spend a romantic night together
HALV	Pink Flamingos	60.2s	A lady licks and eats dog faeces
	Black Swan	62.2s	A lady notices paranormal activity
	Psyco	76.2s	Lady gets killed by intruder in her bathtub
HAHV	Ace-Ventura: Pet Detective	102.1s	Ace Ventura successfully hides his pets
	The Gods Must be Crazy II	67.1s	A couple stranded in the desert steal ostrich eggs
	When Harry Met Sally	100.2s	Sally shows Harry how women fake orgasms at a restaurant

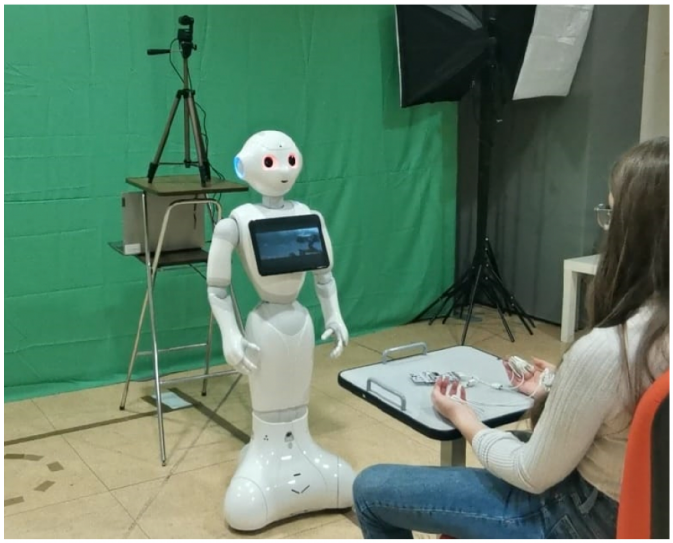

Fig. 1. Example of the experimental setting

(LALV, LAHV, HALV, HAHV), with the respective duration in seconds.

C. The Procedure

Firstly, participants were asked to read the participant information form and sign the consent form. Then, we collected participants' demographics (e.g., age and gender), and their responses to the Empathy Quotient test (EQ) for adults [33]. The EQ test consists of 40 questions (e.g., "It upsets me to see an animal suffer", "I understand if someone is hiding his true emotions") to be rated on a scale from 4 "Absolutely agree" to 1 "Absolutely disagree".

Participants were accommodated on a chair with a small table where they could rest their hands equipped with ECG electrodes (electrode with a negative and a neutral charge on the right palm and electrode with a positive charge on the left palm), and GSR electrodes (electrode with positive and negative charge respectively on the forefinger and middle fingers). The robot was positioned in front of the participants (see Figure 1).

Each participant watched the selected videos in a randomized order. For each video, one of the following conditions was selected in a random order:

- In the *Neutral* condition, the robot stays in a static position without using any non-verbal cues.
- In the *Coherent* condition, the robot displays non-verbal gestures selected in the same class of the video.
- In the *Incoherent* condition, the robot displays non-verbal gestures selected in the opposite class of the video (for example, in the case of a video in LALV class the robot gestures are selected in HAHV class).

The study lasted around 30 minutes on average.

D. Collecting Participants' Multi-modal Signals

To collect the biosignals of the participants, we used BITalino (r)evolution [34], which supports several other sensors, such as Electrodermal Activity (EDA), Electrocardiography, Electromyography, Electroencephalography, Light (LUX), Pushbutton (BTN) and the Accelerometer (ACC). In this work, we only used ECG and EDA signals. We collect 100 samples for a second for each type of signal.

Moreover, participants were recorded with a camera that was located at the back of the robot to get their facial expressions (see Figure 1).

IV. RESULTS

A. The Participants

We recruited 30 participants (19 male, 11 female) aged between 20 and 60 years old (avg. 44.06, st.dv. 12.25). Participants were recruited between staff and students at the University of Naples, Italy. The study was performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki, and participants also gave written informed consent prior to their participation in the study.

Participants were evaluated for their ability to feel empathy (i.e., what others may feel or think) using the Empathy Quotient test. This test is composed of 40 statements on empathy, where questions have a score from 1 to 2 points, for a total score of 80 points and a threshold score lower than 30. To clearly distinguish participants' levels of empathy, we defined four empathy bands based on the scores obtained:

- Low: subjects with a score from 0 to 20;
- Medium-Low: subjects with a score from 21 to 40;
- Medium-High: subjects with a score from 41 to 60;
- High: subjects with a score from 61 to 80.

The majority of participants (94%) fall in both Medium bands (Medium-Low and Medium-High) with equal distribution. Only 6% scored High, and no participants fell in the Low band. This means that the majority of participants have an average level of empathy towards others' feelings and desires.

TABLE IV
EVALUATION OF THE FUSION OF FACE, ECG AND GSR USING THE AMIGOS DATASET AND A GREEDY ALGORITHM.

	Fusion			
	LALV(0)	LAHV(1)	HALV(2)	HAHV(3)
Precision	0.49	0.66	0.85	0.91
Recall	0.91	0.62	0.58	0.47
F1 Score	0.64	0.64	0.69	0.62
Accuracy	0.74	0.83	0.87	0.86

TABLE V
CORRECTNESS OF THE CLASSIFICATION WITH RESPECT TO THE ROBOT'S BEHAVIOURS.

	Robot's behaviours		
Class Emotion	Neutral	Coherent	Incoherent
LALV(0)	100%	97%	77%
LAHV(1)	77%	70%	40%
HALV(2)	77%	70%	27%
HAHV(3)	67%	63%	10%
Average	75%	74%	43%

B. Multimodal Fusion

Since the data are collected continuously, we performed a sampling for each subject that allows us to divide the data related to each video. For the evaluation of the face, the recordings of participants' facial expressions were divided into 12 sub-videos to clearly distinguish each video shown on the robot's tablet, and by eliminating any non-relevant timeframes (i.e., the time to transition from one video to another). We then extracted the frames for each sub-video and evaluated them using the face classifier. In particular, each frame is classified individually and together with the classification associated with the whole video, which is given by the average of the classification probabilities obtained from all the frames.

Biometrical signals were recorded associated with their timestamp and were subjected to an initial screening to remove signals from non-relevant transitional time-frames. We then applied the previously described pre-processing process to the ECG and GSR signals collected to extract their characteristics for each video. At the end of this phase, we obtained the classification results of each video shown for each participant.

As previously described, we evaluated the multimedia classifier with a greedy approach. Results in Table IV show an overall good performance in classification.

C. Evaluation of Robot's Behaviours

In this study, we were not interested in the classification performance *per se*, but in evaluating the impact of the three behaviours of the robot (Neutral, Coherent and Incoherent) on participants. To this extent, we evaluated the correctness of the classification based on the robot's behaviours. Table V shows the percentage of correctness obtained by videos belonging to a certain class. As we can observe from the table, Neutral behaviour has a higher percentage of correctness, and it represents the baseline to evaluate the robot's contribution in emotion elicitation. Classification results for Neutral behaviours differ slightly from the percentage obtained with Coherent behaviour. As we foresee, a significantly lower percentage of correct results is obtained in the case of Incoherent behaviour, meaning that the emotion elicited in some of the users was different with respect to the one elicited by the video displayed on the tablet. This result is particularly evident for classes characterized by high arousal (HALV and HAHV).

To evaluate if such a lower result corresponds to the effect of the robot's behaviour in changing the elicited emotion,

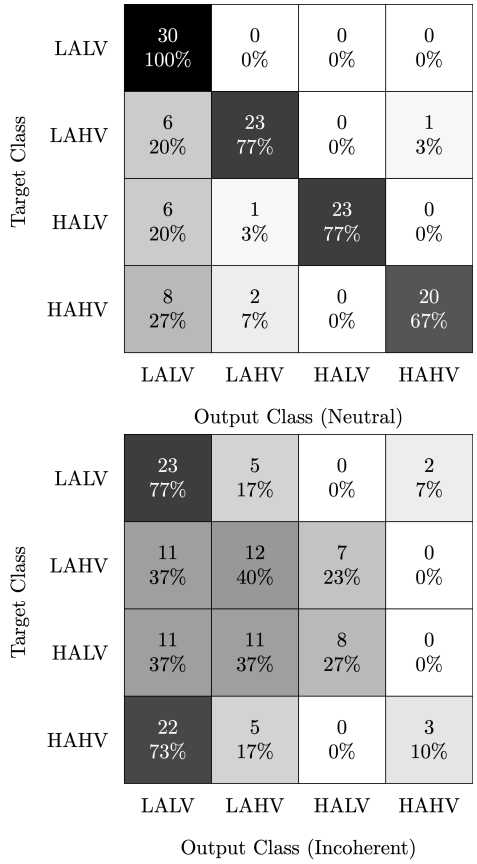

Fig. 2. Confusion matrix for the Neutral (top image) and the Incoherent (bottom image) conditions

we, then, evaluated the confusion matrices obtained for the Neutral and Incoherent conditions (see Figure 2). In general, as shown in the Neutral condition, errors of the multi-modal classifier are towards a miss-classification of the class LALV. However, as shown in the Incoherent case, there is an increase of classification of the perceived emotion towards the class opposite to the one related to the video (and in line with the behaviour of the robot). These results are in favour of the possible ability for the robot in shifting the elicited emotion. This is particularly relevant in the case of the class HAHV(3) where the majority of the classifications, in the case of an Incoherent robot behaviours, lie in the opposite class (e.g., LALV(0). Smaller effects are on the pair LAHV(1)-HALV(2).

TABLE VI
CORRECTNESS EVALUATION OF THE OPPOSITE CLASSES FOR THE INCOHERENT ROBOT'S BEHAVIOUR.

Class Emotion	Prob(0)	Prob(1)	Prob(2)	Prob(3)
LALV(0)	0.820	**0.268**	0.303	0.316
LAHV(1)	0.448	0.598	**0.322**	0.297
HALV(2)	0.427	**0.311**	**0.623**	0.270
HAHV(3)	**0.545**	0.285	0.289	**0.541**

TABLE VII
POST-DOC ANALYSIS WITH BONFERRONI CORRECTION FOR HALV(2) AND HAHV(3) CLASSES.

HALV(2) class		
Difference of the mean	Prob(1)	Prob(2)
Neutral - Incoherent	**-0.249**	**0.258**
Coherent - Incoherent	**-0.206**	**0.246**
HAHV(3) class		
Difference of the mean	Prob(0)	Prob(3)
Neutral - Incoherent	**-0.297**	**0.302**
Coherent - Incoherent	**-0.247**	**0.328**

To further investigate the differences in the classification according to the robot behaviours, we conducted an ANOVA statistical analysis on the probabilities returned by the multimodal classification process. In our case, we considered as a factor the robot's behaviour (Neutral, Coherent and Incoherent), while the dependent variable is the classification obtained for a given class, applying the Bonferroni correction for multiple comparisons. Table VI shows the results obtained from this analysis, where the rows represent the real classes, while Prob(0), Prob(1), Prob(2) Prob(3) indicate respectively the obtained probability of belonging to class LALV(0), LAHV(1), HALV(2) and HAHV(3). The elements highlighted in the table indicate significant differences.

Considering the different behaviours for the HALV(2) class, we obtained a significant difference both for Prob(2), which represents the probability of belonging to the HALV(2) class (real class) and for Prob(1), which represents the probability of belonging to the LAHV(1) class (i.e. the opposite class). We can observe similar results for the HAHV(3) class. This means that the robot's behaviours for these two classes had an impact on the emotional reaction to the video, and in many cases managed to shift participants' emotions to the opposite class through Incoherent Behaviour. This is also confirmed by the Bonferroni correction with significant difference on the means around 20-30% (see Table VII).

Considering the LAHV(1) class, we found a significant difference only in the Prob(2) case, which represents the probability of belonging to the opposite class (HALV(2)), but not in Prob(1). Finally, for the class LALV(0) we have identified a significant difference with the case Prob(1), which does not correspond to the opposite class.

V. CONCLUSIONS

In this work, we presented a multi-modal architecture for classifying emotions using participants' facial expressions and two different biosignals (electrocardiogram and the galvanic skin response). We considered the emotions in terms of valance and arousal, and we defined four classes of emotions, which refer to the four quadrants of Russell's Circumplex model (LALV, LAHV, HALV, HAHV).

We tested our model in a user study with a physical robot to understand the impact of the robot's emotional non-verbal behaviour on the emotion elicitation process. To this extent, we designed three robot behaviours (incoherent, coherent and neutral) and we used them to manipulate people's emotional reactions while watching short clips extracted from a standardized dataset used for emotion elicitation.

Our results showed that there are significant differences between HALV(2) and HAHV(3) classes. In particular, the non-verbal behaviour modelled for the robot was effective for the classes and, in the case of incoherent behaviour, it could change people's emotions of the considered class to the opposite one. However, we did not observe a similar effect for the LALV(0) and LAHV(1) classes. Our main hypothesis to explain our results is that in the specific setting, the robot behaviours are able to get the attention from the video and so change the valence to the opposite class but without being able to achieve high arousal. Indeed, eliciting an emotional reaction with high arousal could be challenging by the use only of accompanying emotional gestures and need additional interaction strategies.

These results highlight the importance of social robots as a supportive role in healthcare scenarios where they could help vulnerable people, such as older people and patients, to cope with an emotional situation that may negatively affect their quality of life. In future work, we aim to further investigate the robot's eliciting behaviours using this framework and by considering not only the effect on the opposite class. Moreover, different emotional interaction strategies have to be explored to investigate how to impact arousal with different robots.

ETHICAL IMPACT STATEMENT

a) Potential negative applications: The attempt of robots in mimicking human ability to feel and show emotions to elicit a change in people's affective responses may be perceived as a robot's deception. Robot deception presents a philosophical and psychological controversy where a person may not be willing to further use the robot in the future, or they may be affected by a sense of betrayal or feel less confident in their own abilities and judgement.

b) Generalizability and Potential Bias: The majority of participants were white, and recruited within the University's premises, therefore, our results may not be generalised to participants of other nationalities, backgrounds or different levels of empathy. Future works should include data representing a larger and more inclusive population.

c) Privacy: Participants received a brief of the human-robot interaction prior to taking part in the study. They were aware of the data (face, GSR and ECG) collected and provided written consent. Data were securely stored in the research institute, and accessible only to the experimenter and researchers of the project. The data has been destroyed after the analysis.

REFERENCES

[1] M. Shao, M. Snyder, G. Nejat, and B. Benhabib, "User affect elicitation with a socially emotional robot," *Robotics*, vol. 9, no. 2, 2020. [Online]. Available: https://www.mdpi.com/2218-6581/9/2/44

[2] S. Rossi, M. Larafa, and M. Ruocco, "Emotional and behavioural distraction by a social robot for children anxiety reduction during vaccination," *International Journal of Social Robotics*, vol. 12, no. 3, pp. 765–777, 2020. [Online]. Available: https://doi.org/10.1007/s12369-019-00616-w

[3] S. Rossi, S. J. Santini, D. Di Genova, G. Maggi, A. Verrotti, G. Farello, R. Romualdi, A. Alisi, A. E. Tozzi, and C. Balsano, "Using the social robot nao for emotional support to children at a pediatric emergency department: Randomized clinical trial," *J Med Internet Res*, vol. 24, no. 1, p. e29656, Jan 2022.

[4] J. Xu, J. Broekens, K. Hindriks, and M. A. Neerincx, "Effects of bodily mood expression of a robotic teacher on students," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014, pp. 2614–2620.

[5] I. Torre, E. Carrigan, R. McDonnell, K. Domijan, K. McCabe, and N. Harte, "The effect of multimodal emotional expression and agent appearance on trust in human-agent interaction," in *Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games*, ser. MIG '19. New York, NY, USA: Association for Computing Machinery, 2019.

[6] M. Spezialetti, G. Placidi, and S. Rossi, "Emotion recognition for human-robot interaction: Recent advances and future perspectives," *Frontiers in Robotics and AI*, vol. 7, 2020.

[7] M. Soleymani, G. Chanel, J. J. M. Kierkels, and T. Pun, "Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses," in *2008 Tenth IEEE International Symposium on Multimedia*, 2008, pp. 228–235.

[8] F.-F. Kuo, M.-F. Chiang, M.-K. Shan, and S.-Y. Lee, "Emotion-based music recommendation by association discovery from film music," in *Proceedings of the 13th Annual ACM International Conference on Multimedia*, ser. MULTIMEDIA '05. New York, NY, USA: Association for Computing Machinery, 2005, p. 507–510.

[9] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 211–223, 2012.

[10] L. Fiorini, F. G. C. Loizzo, G. D'Onofrio, A. Sorrentino, F. Ciccone, S. Russo, F. Giuliani, D. Sancarlo, and F. Cavallo, "Can i feel you? recognizing human's emotions during human-robot interaction," in *Social Robotics*. Cham: Springer Nature Switzerland, 2022, pp. 511–521.

[11] A. Lim, T. Ogata, and H. Okuno, "Towards expressive musical robots: A cross-modal framework for emotional gesture, voice and music," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2012, 12 2012.

[12] N. Riether, F. Hegel, B. Wrede, and G. Horstmann, "Social facilitation with social robots?" in *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 41–48. [Online]. Available: https://doi.org/10.1145/2157689.2157697

[13] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Information Fusion*, vol. 59, pp. 103–126, 2020.

[14] S. Rossi, T. Cimmino, M. Matarese, and M. Raiano, "Coherent and incoherent robot emotional behavior for humorous and engaging recommendations," in *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2019, pp. 1–6.

[15] S. Fan, Z. Shen, B. L. Koenig, T. Ng, and M. S. Kankanhalli, "When and why static images are more effective than videos," *IEEE Transactions on Affective Computing*, vol. 14, no. 01, pp. 308–320, jan 2023.

[16] K. Rymarczyk, L. Zurawski, K. Jankowiak-Siuda, and I. Szatkowska, "Do dynamic compared to static facial expressions of happiness and anger reveal enhanced facial mimicry?" *PLOS ONE*, vol. 11, no. 7, pp. 1–15, 07 2016.

[17] C. Frith, "Role of facial expressions in social interactions," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, pp. 3453–3458, 2009.

[18] M. Rescigno, M. Spezialetti, and S. Rossi, "Personalized models for facial emotion recognition through transfer learning," *Multimedia Tools and Applications*, vol. 79, no. 47, pp. 35 811–35 828, 2020.

[19] Z. Yu, X. Li, and G. Zhao, "Facial-video-based physiological signal measurement: Recent advances and affective applications," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 50–58, 2021.

[20] D. Nikolova, P. Petkova, A. Manolova, and P. Georgieva, "Ecg-based emotion recognition: Overview of methods and applications," *ANNA'18; Advances in Neural Networks and Applications 2018*, pp. 1–5, 2018.

[21] L. Santamaria-Granados, M. Muñoz-Organero, G. Ramírez-González, E. W. Abdulhay, and N. Arunkumar, "Using deep convolutional neural network for emotion detection on a physiological signals dataset (amigos)," *IEEE Access*, vol. 7, pp. 57–67, 2019.

[22] F. Agrafioti, D. Hatzinakos, and A. K. Anderson, "Ecg pattern analysis for emotion detection," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 102–115, 2012.

[23] M. A. Hasnul, N. A. A. Aziz, S. Alelyani, M. Mohana, and A. A. Aziz, "Electrocardiogram-based emotion recognition systems and their applications in healthcare—a review," *Sensors*, vol. 21, no. 15, p. 5015, 2021.

[24] G. Wu, G. Liu, and M. Hao, "The analysis of emotion recognition from gsr based on pso," in *2010 International Symposium on Intelligence Information Processing and Trusted Computing*, 2010, pp. 360–363.

[25] N. Ahmed, Z. A. Aghbari, and S. Girija, "A systematic survey on multimodal emotion recognition using learning algorithms," *Intelligent Systems with Applications*, vol. 17, p. 200171, 2023.

[26] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 479–493, 2021.

[27] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[28] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the royal statistical society. series c (applied statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, p. 84–90, may 2017.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[31] M. Marmpena, A. Lim, and T. S. Dahl, "How does the robot feel? perception of valence and arousal in emotional body language," *Paladyn, Journal of Behavioral Robotics*, vol. 9, no. 1, pp. 168–182, 2018. [Online]. Available: https://doi.org/10.1515/pjbr-2018-0012

[32] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, "Decaf: Meg-based multimodal database for decoding affective physiological responses," *IEEE Transactions on Affective Computing*, vol. 6, no. 3, pp. 209–222, 2015.

[33] E. J. Lawrence, P. Shaw, D. Baker, S. Baron-Cohen, and A. S. David, "Measuring empathy: reliability and validity of the empathy quotient," *Psychological Medicine*, vol. 34, no. 5, p. 911–920, 2004.

[34] H. Plácido da Silva, J. Guerreiro, A. Lourenco, A. Fred, and R. Martins, "Bitalino: A novel hardware framework for physiological computing," 01 2014.