IEEE Transactions on Cybernetics Manuscript (This is the author's version of an article that has been published. Changes were made by the publisher prior to publication. (https://doi.org/10.1109/TCYB.2020.2974688)

A Multimodal Emotional Human-Robot Interaction Architecture for Social Robots Engaged in Bi-directional Communication

Alexander Hong, Nolan Lunscher, Tianhao Hu, Yuma Tsuboi, Xinyi Zhang, *Student Member*, *IEEE*, Silas Alves, *Member*, *IEEE*, Goldie Nejat, *Member*, *IEEE*, and Beno Benhabib

***Abstract*— For social robots to effectively engage in Human-Robot Interaction (HRI) they need to be able to interpret human affective cues and to respond appropriately via display of their own emotional behavior. In this paper, we present a novel multimodal emotional HRI architecture to promote natural and engaging bi-directional emotional communications between a social robot and a human user. User affect is detected using a unique combination of body language and vocal intonation, and multi-modal classification is performed using a Bayesian Network. The Emotionally Expressive Robot utilizes the user's affect to determine its own emotional behavior via an innovative two-layer emotional model consisting of deliberative (hidden Markov Model) and reactive (rule-based) layers.**

The proposed architecture has been implemented via a small humanoid robot to perform diet and fitness counselling during HRI. In order to evaluate the Emotionally Expressive Robot's effectiveness, a Neutral Robot that can detect user affects, but lacks emotional display, was also developed. A between-subjects HRI experiment was conducted with both types of robots. Extensive results have shown that both robots can effectively detect user affect during real-time HRI. However, the Emotionally Expressive Robot can appropriately determine its own emotional response based on the situation at hand and, therefore, induce more user positive valence and less negative arousal than the Neutral Robot.

***Index Terms*— Human-Robot Interaction, Multimodal Affect Recognition, Robot Emotion Model, Social Robots**

I. Introduction

EMOTIONS are an essential part of human behavior that can influence how people communicate and make decisions [1]. In order for social robots to effectively interact with people and provide assistance with a number of everyday tasks, they need to intelligently recognize and classify human affective states and, in turn, respond appropriately using their own emotional assistive behaviors. Such interactions are more engaging and can enhance a robot's ability to assist people [2]. Emotions are also important for long-term interactions, in order to maintain social-emotional relationships [3]. Such long-term interactions are imperative for users to obtain the benefits of interacting with robots during daily activities [4], and in turn for these robots to have impact on people's lives in the near future [5]. Furthermore, it is important to consider the variety of cultural attitudes towards robots [6] and how they influence affective interactions, as people from diverse cultures may interpret robot emotions differently [7]. Therefore, emotion recognition is a key challenge for enabling robots to communicate with people [8].

Human affective cues can be inferred from natural communication modalities, such as facial expressions [9-18], vocal intonation [19-21], and body language [22-26]. These modes correlate strongly to a person's affective state [27]. The use of multimodal affect recognition systems, which combine two or more modalities, has also been investigated for human-computer interactions (HCI) [28] and human-robot interactions (HRI) [2]. Multimodal inputs provide two key advantages [29]: a combination of modalities provides diverse complementary information, which increases robustness and performance; and, when one modality is unavailable, due to occlusion and/or noise, a multimodal recognition system can use the remaining modalities to estimate affect.

Once the robot has estimated a user's affect, it can use this information to determine its own emotional behavior via a (robot) emotional model. The incorporation of such a model would allow a social robot to create more natural and engaging interactions with users [24, 30]. It would also allow a robot to respond to different situations that can vary from light-hearted interactions to more serious interactions.

The ability to both detect and express emotions promotes bi-directional interactions and, thus, allows robots to establish social relationships with people [31]. A two-layer emotional model can be utilized to improve the adaptive performance of a robot [24, 32]. Such a model allows the robot to determine its appropriate emotional behaviors based on user input and the current HRI scenario, as well as react to unexpected or potentially unsafe situations that may occur during HRI. In general, social robots communicate their emotions through the use of different modalities, including facial expressions [9, 33], vocal intonation [34], and gestures and body language [32], as well as by using a combination of these modes to create multimodal emotional displays [31, 38-40].

In general, social robotics research has either focused on

This work was supported by the Natural Sciences and Engineering Research Council of Canada, the Canada Research Chairs program, the Canadian Consortium on Neurodegeneration in Aging, and AGE-WELL.

The authors are with the Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, Canada (emails: {alex.hong, nolan.lunscher, tianhao.hu, yuma.tsuboi}@mail.utoronto.ca, {xinyiz.zhang,silas.alves}@utoronto.ca,{nejat, benhabib}@mie.utoronto.ca).

multi-modal user affect detection [41-45] or on robot emotional models [24, 30, 32, 46-50]. Only a few have considered the use of multi-modal user affect to determine a robot's multimodal emotional expressions using a robot emotional model [31, 36-38]. However, the emotional models used by these robots do not consider unexpected situations that may occur during HRI. Developing robots that can detect human affect and respond with emotional expression themselves has many advantages. For instance, it allows a robot to directly focus on a person's wellbeing [24] and promotes cooperation from potential users [51].

Herein, we propose a novel multimodal emotional HRI architecture. It utilizes a multimodal affect recognition and classification system, as well as a robot emotional model that considers the user's affective intent and the interaction scenario in order for a social robot to determine its own emotional behavior. Our main contributions are on the development of: (1) the multimodal affect recognition and classification system to determine user affect that uniquely considers the combination of body language and vocal intonation, modalities that have not yet been explored together for social human-robot interaction; and, (2) a two-layer robot emotional model that uniquely considers unpredictable robot emotion expressions. The two-layer robot model uses a Hidden Markov model (HMM) for the deliberative layer and a rule-based approach for the reactive layer.

The proposed architecture was implemented and evaluated on a Nao humanoid robot platform. The robot engaged users in multimodal bi-directional HRI for the purpose of diet and fitness planning. To the authors' knowledge, Autom [52] is the only other robot designed for this particular application, however interactions are purely through its tablet, and therefore it is unable to perform bi-directional affective HRI.

II. Related Work

Social robots with either automated multimodal affect recognition systems have been reported in [43-45, 53-56], or with their own emotional response models have been presented in [9, 10, 24, 30, 32, 46-49]. However, only a few robots use multimodal human affect detection to determine their own emotional behavior during HRI [31, 36-38].

A. Robot Emotional Models

Robots with emotions can improve the overall interactive experience with users, as they tend to be perceived as more intelligent and natural to interact with [30]. Robot emotional models that determine these emotions have been classified into single-layer models, which consist of only deliberative emotions [9, 10, 30, 47, 48, 57], or two-layer models, which consider both deliberative and reactive emotions [29, 32, 49].

1) Single-Layer Emotional Models

In [30], the Kismet robot used an emotion-inspired model to create life-like behaviors. Its emotion model used a continuous 3D affect space comprising arousal, valence and stance. Kismet could recognize a user's affective communications through tone of voice by classifying prosodic patterns. Its own affect would be communicated through a combination of facial expressions, body posture, and vocal intonation.

In [47], the RWI B21 robot, with a virtual face, was used as a receptionist, displaying emotional responses to human visitors. The robot incorporated a set of short-lived basic emotions, longer lasting moods that made the robot feel positive or negative, as well as an attitude that was represented as a mood level and a familiarity rating associated with a specific person. Robot emotions were displayed via facial expressions, while mood was displayed as posture changes.

In [48], an emotional model based on Gross' process model of emotion regulation was developed for a robot with a virtual face. The model consisted of an HMM using hidden emotional states and observable expression states. The observable states were presented as facial expressions with varying intensities of emotion. A personality suppression of expression mechanism was added to the model, allowing for the robot's observable states to be controlled by manipulating the likelihoods of higher or lower intensity expressions occurring.

In [10], a robot used an Ortony, Clore and Colins (OCC) [58] -based emotion model. The robot updated the transition probabilities of the model through an active field state space, which modeled influences as field forces in the affective dimensions of valence, arousal and stance. The transition probabilities of the model were updated based on the user's affect, determined through facial expressions. The robot's emotional expressions were, then, displayed through facial expressions and body language.

In [9], a human-like robotic head detected a user's affect based on facial expressions and, in turn, displayed its own facial expressions based on its mood. The robotic head used the Circumplex Model [59] to represent its mood in terms of valence and arousal, and the Five Factor Model of personality to define its personality. User affect was transferred into robot mood variables using a linear relationship. Mood was, then, processed by a Fuzzy Kohonen clustering network for the robot to display via facial expressions.

2) Multi-Layer Emotional Models

In [24], a human-like robot, Brian, used a robot emotion architecture that introduced the concept of deliberative and reactive emotion layers. The deliberative layer used an online learning-based Markov model to determine the robot's future emotional state based on the user's affect (determined through body language), the robot's drives, and the robot's emotional state. The reactive layer was proposed for interactions that require an immediate response from the robot to ensure safety (e.g., when a user enters the workspace of the robot, Brian would trigger a reactive fear emotion and stop movements to prevent itself from hitting the user with its arms). This paper focused mainly on the design and implementation of the deliberative layer to determine the emotional behaviors of the robot during a user schedule making activity.

In [32], a simulated mobile robot used emotions to improve its navigation performance in unknown environments. The robot deliberative layer employed a A* method to generate a path using an occupancy grid map. This layer used deliberative

emotions to modify the type of path being generated by changing the parameters of the A* cost function. A reactive controller was used to generate the robot's heading and speed to follow the planned path. This controller used reactive emotions to modulate the directional and speed control parameters and to trigger re-planning if required.

In [49], the robot KaMERo used a reactive-deliberative emotion architecture. The reactive emotion module generated immediate emotional responses to stimuli. The deliberative emotion module used the OCC model to appraise an external event to generate an emotional response. The Big-Five personality model was used to define the robot's personality factors, which were integrated into the emotion- generation function to change the maximum intensity, rising/decaying slopes, and duration of the generated emotions.

B. Multi-modal HRI

The multimodal affect recognition systems that have been used for HRI have mainly utilized the following primary modes: facial and vocal expressions [41, 54-56, 60, 61], facial and gait [45], facial and eye gaze [53], or a combination of physiological signals, eye gaze, and hand gestures [44, 62]. However, the robots using these systems were not capable of displaying their own multimodal affect and, hence, did not engage in a bi-directional affective social interaction with the user. To-date, only a handful of robots that engage in multimodal emotional HRI, where both the user and robot engage in emotional behavior, have been the focus of recent research. These robots can be classified into those that mimic the emotions of users [36, 63], and those that determine their own emotions based on user affect [31, 37, 38].

In [36], a humanoid robotic head, Muecas, was used to mimic affect using a user's facial and vocal expressions as inputs. Facial expression recognition was achieved through Gabor filtering and dynamic Bayesian network classification. Vocal features of speech rate, pitch, and energy were extracted and fed to a dynamic Bayesian network classifier to determine vocal expressions. A multimodal decision-level fusion, also based on a final dynamic Bayesian network, combined the two modalities to recognize happiness, sadness, anger, neutral, and fear. Experiments showed that Muecas could mimic the affective states through both facial expressions and body language using Action Unit (AU) reconstruction model.

In [63], the Nao humanoid robot was used to recognize a user's emotion, *neutral*, *happy*, *anger*, *surprise*, *fear*, *disgust* or *sadness*, from his/her vocal intonation, facial expression and gestures. Vocal features such as fundamental frequency, and Mel-frequency Cepstral Coefficients were extracted using correlation analysis. Facial expression features were extracted using a Gabor filter, and analyzed using an Extreme Learning Machine (ELM) model to recognize the user's facial expression. Gesture recognition was achieved using the Angular Metric for Shape Similarity (AMSS) algorithm to calculate similarity between the gesture template and the gesture tracked in real time. All three modalities were classified using a Naïve Bayesian classifier, where the posterior was considered as the detected emotion. Finally, the robot was able to display its own emotional states by mimicking the user's emotion through gestures and eye colors.

In [31], a stuffed-animal-like robot, CuDDler, used vision and sound to detect the affective states. Interaction sounds were detected using sound signatures. Local binary pattern features were extracted from facial information and input into a linear support vector machine (SVM) classifier to determine facial affect. Experiments showed that the robot was able to recognize the affective acts of pat, hit, and stroke, and responded appropriately to these situations by blinking its eyes, and displaying gestures and sounds.

In [37], the humanoid robot AMI could classify user facial and vocal expressions into the affective states of *happy*, *sad*, and *angry*. Facial AU features of lips, eyebrows, and the forehead, and vocal features of phonetic and prosodic properties were extracted and classified as affective states using neural networks. Multimodal affect recognition was accomplished using decision-level fusion based on weighted sum. The robot used this information to produce its own affect through a synthesizer that considered emotional drives, human emotional status, and a decaying term that restored the robot's emotional status to neutral. It displayed its own appropriate affect through dialogue, facial, and gesture expressions.

In [38], a mobile robot determined a user's *neutral*, *happy*, *sad*, *fear*, or *anger* affective states by recognizing facial and vocal expressions. Facial AUs from the eyes, eyebrows, nose, and mouth were extracted using principal component analysis. Voice features of sampling frequency, pitch, and volume level were extracted using the Praat vocal toolkit. Both feature vectors, and the robot's social profile (sympathetic, anti-sympathetic, and humorous) were input into a Bayesian network that determined the robot's affect, which it displayed using facial and vocal expressions. Experiments consisted of participants rating the robot's response as *funny*, *neutral*, or *aggressive* based on its social profile.

C. Summary of Related Work

Body language plays an important role in conveying changes in human emotions [64], as people communicate their affect through a dynamic range of body motions [42, 65]. Within our system, we consider both static and dynamic body language. Static body language is an important source of information for affect expression [66], as changes in static body language can induce changes in affective states [67]; while dynamic properties of emotional expressions (e.g., time and vigor) can represent emotional intensity [68]. Additionally, it has also been shown that vocal utterances strongly correlate to body language [42]. Vocal intonation plays an important role in conveying changes in emotion through pitch, tempo, and loudness [69].

Current multimodal affect recognition systems for HRI have mainly used facial expressions and vocal intonation [2, 35, 37, 38]. Although the face is an important source of emotional information, facial expressions can be more easily controlled and faked compared to body expressions [70, 71]. However, body language has been shown to provide more distinctive cues than facial expressions especially when discriminating between

positive and negative emotions [70, 72]. Vocal intonation is also regarded to be more expressive and less controlled than the face [73], since the effort of controlling the voice can lead to either being overcontrolled and unnatural or totally lacking in control (i.e., emotion leakage) [74].

Our proposed HRI architecture incorporates a multimodal human affect recognition and classification system that uniquely utilizes both these important modes to determine user affect in order for a social robot to respond appropriately during HRI using its own emotional behavior.

With respect to the existing robot emotional models, the majority have only a deliberative layer. The concept of using a two-layer emotional model that also considers reactive emotions has been proposed [24, 32, 49], but has not been implemented for social HRI. Furthermore, the proposed deliberative layers work separately from the reactive layers.

Our proposed two-layer deliberative and reactive robot emotional model determines the robot's emotions by uniquely considering the uncertainty in HRI scenarios. Namely, we introduce a novel emotion observation feedback system that allows observations of the emotional expressions that the robot physically implements during HRI to impact how its deliberative emotions are chosen. The reactive layer allows the robot to respond to unexpected or unsafe situations during interactions, while the deliberative layer is used to determine the robot's emotions based on the task at hand, the user's affect and the robot's own previous emotional expression.

III. Proposed Multimodal HRI Architecture

Our proposed multimodal emotional HRI system architecture, Fig. 1, comprises three main subsystems: the multimodal affect recognition sub-system (MARS), the robot emotion model (REM), and the interaction activity sub-system (IAS). 3D data from the Kinect sensor and vocal signals from the microphone are used by the MARS to first determine and classify body language and vocal intonation each in terms of valence and arousal. These outputs are combined into an affect vector for multimodal affect classification of the user. The REM determines the robot's emotions based on the deliberative or reactive layer. The deliberative layer uses the classified affect of the user, the robot's desires and drives, and the robot's previously displayed emotion to determine the robot's current deliberative emotion. Whereas, the reactive layer uses input from the robot's touch sensors and 2D camera to determine the robot's reactive emotion. The emotion with the highest priority is, then, chosen to be expressed by the robot via multi-modal outputs using the robot's low-level controller. Observations of the emotional expression are used as feedback to the deliberative layer to identify whether deliberative or reactive emotional expressions were successfully implemented. The IAS is used to determine the robot's appropriate behavior based on the activity at hand. This behavior is displayed using the chosen emotion. IAS also provides user compliance and activity progression for REM to verify whether the robot's desires and drives are met. Each of the modules is discussed in more details below.

A. Multimodal Affect Recognition System (MARS)

MARS is used to classify a person's affect based on a 2D valence-arousal scale. Valence is used to define a user's level of pleasure, and arousal is used to define his/her excitation level [59]. We chose the valence-arousal scale as it encompasses all possible affective states and their variations [75]. In addition, valence and arousal better represent experimental and clinical findings compared to categorical emotional models (e.g., happy, angry, sad) [76]. We utilize decision-level fusion to effectively estimate a user's multimodal affect during HRI based on both body language and vocal intonation. Namely, affect from each of the two modalities is determined first and, then, combined to determine overall affect.

1) Body Language

Postures and body movements have been shown to be directly correlated to a person's affect [42, 65] and be used effectively to communicate affect during social interactions [2]. Body language has been defined as an interaction of at least two seconds long [42]. Our previous work focused on identifying body language features and validating these features for automated recognition and classification by a robot [77, 78]. Herein, we utilize these for the body language mode for our multimodal affect classification. Namely, we utilize the body language descriptors of bowing/stretching trunk, opening/closing of the arms, vertical head position and motion

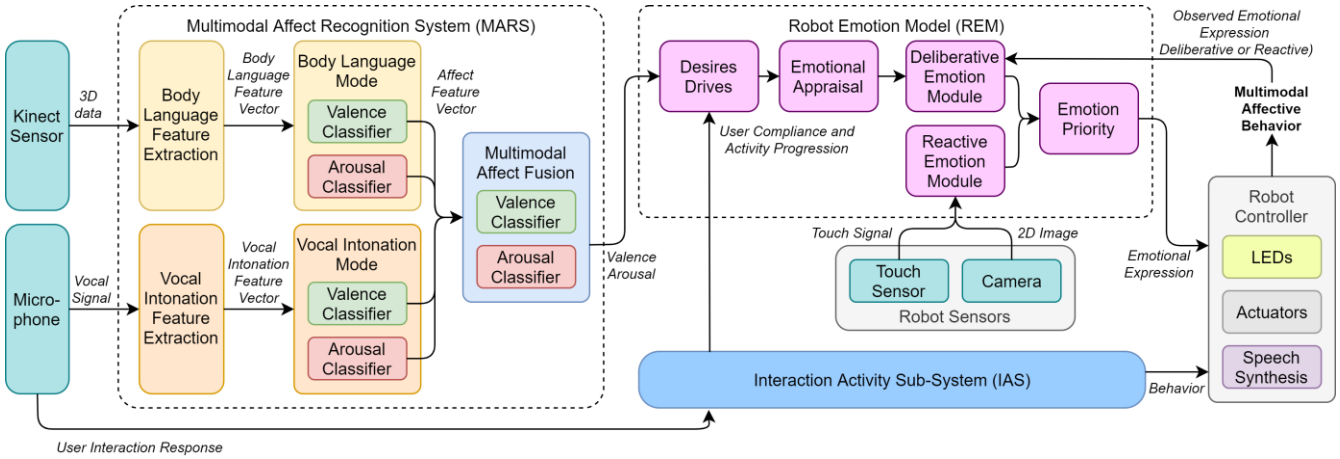

Fig. 1. Proposed Multimodal Emotional HRI Architecture.

of the body, forward/backwards head position and motion of the body, expansiveness of the body, and speed of the body. The details of the features and their descriptors are provided in Table A in the supplementary material.

Real-time identification and tracking of these features are achieved by using 3D information of the user's body provided by a Kinect™ 3D sensor. Namely, 20 position coordinates of the body are identified and tracked using the Kinect Skeleton [79], including on the head, shoulder center, spine, hip center, both left and right hands, wrists, elbows, shoulders, hips, knees, ankles, and feet. Dynamic body language features are, then, calculated using the tracked points, forming a feature vector at a sampling rate of 30 frames per second. Once the feature vector is obtained, affect classification takes place. Random forest decision trees are used to classify both valence and arousal.

2) Vocal Intonation

Audio signal patterns in vocal intonation, exclusive of speech content, are used to estimate vocal intonation of the user. Vocal intonation features are identified based on the local extrema and flatness in amplitude of vocal signals [80]. These are based on the peaks and plateaus of the signal. In order to extract vocal intonation features in real-time HRI scenarios, a noise-cancelling microphone array is needed, such as the XMOS Vocalfusion Speaker Circular Array (XVF3100) used herein. These features are directly extracted from the vocal signals using the Nemesysco QA5 SDK software [81]. The extracted features and their descriptors are provided in Table B in the supplementary material. Vocal intonation is, then, classified as valence and arousal using model trees.

3) Multimodal Affect Fusion

Decision-level fusion combines the classified valence, v_b, and arousal, a_b, values from body language, and the valence, v_v, and arousal, a_v, values from vocal intonation, forming an affect feature vector for decision-level affect classification. Both body language and vocal intonation affective values have a one-to-one correspondence as they are taken from the same interaction time interval for decision-level fusion.

In order to classify the affect feature vector, a Bayesian network is used for multimodal valence and arousal classification. For the multimodal affect vector, $\mathbf{c}_m = \begin{bmatrix} v_m \\ a_m \end{bmatrix}$, containing multimodal valence value, v_m, and multimodal arousal value, a_m, the joint probability function is defined as:

$$P(\mathbf{c}_m, v_b, a_b, v_v, a_v) = P(v_b \mid \mathbf{c}_m)P(a_b \mid \mathbf{c}_m)P(v_v \mid \mathbf{c}_m)P(a_v \mid \mathbf{c}_m)P(\mathbf{c}_m). \quad (1)$$

From the joint distribution, the posterior probability of $\mathbf{c}_m$ can be obtained by applying Bayes' Theorem:

$$P(\mathbf{c}_m \mid v_b, a_b, v_v, a_v) = \frac{P(v_b \mid \mathbf{c}_m)\, P(a_b \mid \mathbf{c}_m)\, P(v_v \mid \mathbf{c}_m)\, P(a_v \mid \mathbf{c}_m)}{P(v_b, a_b, v_v, a_v)}. \quad (2)$$

The multimodal class, $\mathbf{c}_m$, with the highest probability is chosen as the output of the multimodal affect recognition system, and the corresponding valence and arousal values are passed to REM.

B. Robot Emotional Model (REM)

REM utilizes a user's affect levels (from MARS) and interaction responses as inputs in order to generate appropriate emotional states and expressions for the robot. The model utilizes discrete emotions, with associated dynamic emotional expressions. REM encompasses a novel two-layer model approach comprising deliberative and reactive layers. To the authors' knowledge, this is the first use of such an approach for social robots engaging in HRI. We chose to use this two-layer system as it can provide an adaptive emotional model that can both (*i*) respond to unpredictable situations, and (*ii*) improve user engagement in the activity.

Deliberative emotions are aimed at keeping the user engaged during HRI, and are determined based on the user's affect, the robot's own previous emotional state, and the interaction scenario at hand. Reactive emotions are activity-independent emotions, and are used as emotional responses to stimuli associated with potentially dangerous and unpredictable scenarios during HRI in order to mitigate risk and communicate the robot's safety concerns to the user.

1) Robot Deliberative Emotions

The deliberative emotion layer is the main decision-making layer in charge of determining the emotional state of the robot based on user affect and interaction input, and its own desires and drives. The deliberative emotional model utilizes a hidden Markov model (HMM). An HMM was chosen since it can create human-like emotional agents, through the use of only small sets of discrete emotions [10, 48].

Herein, we implement a novel observation feedback system, which allows observations of the emotional expressions that the robot physically implements during HRI to impact how its deliberative emotions are chosen. Namely, the observations identify whether deliberative or reactive emotional expressions were successfully implemented, and provide this as direct feedback to the deliberative emotion layer.

The deliberative emotional model is represented as:

$$X_{t+1} = A_t X_t, \quad (3)$$

$$Y_{t+1} = B X_{t+1}, \quad (4)$$

where X_{t+1} and X_t represent the robot emotional state vector at time $t+1$ and the current time t, respectively. A_t is the $N \times N$ emotional state transition matrix. Y_{t+1} represents the emotional expression display at time $t+1$, and B is the $M \times N$ block diagonal matrix representing the robot emotional expression probability distribution.

The emotional transition matrix, A_t, describes the emotional state probability distribution, where a transition from state X_t to X_{t+1} occurs with probability:

$$P(X_{t+1} = e_i \mid X_t = e_j) = a_t^{i,j} P(X_t = e_j). \quad (5)$$

where e_i and e_j represent single emotional states in the set of possible emotional states $E = [e_1 \dots e_N]$. B determines the likelihoods of the various emotional expressions occurring. An expression Y_{t+1} occurs with probability:

$$P(Y_{t+1} = ed_i \mid X_{t+1} = e_j) = b_{i,j} P(X_{t+1} = e_j), \quad (6)$$

where ed_i represents a single emotional display in the set of

available expressions $ED = [ed_1 \dots ed_M]$. $a_t^{i,j}$ and $b_{i,j}$ are the elements of matrices A_t and B, respectively, and have the following property:

$$0 \leq a_t^{i,j}, b_{i,j} \leq 1 \text{ and } \sum_{i=1}^{N} a_t^{i,j}, \sum_{i=1}^{M} b_{i,j} = 1. \quad (7)$$

2) Deliberative Emotion Influences

The robot's emotional states are influenced by the robot's previous emotional expression, the user's affect and interaction inputs. The influences of the inputs from the user are determined through an emotional appraisal procedure that incorporates the robot's desires and drives.

Robot desires are outcomes that the robot wishes to occur with respect to the user's affect and behavior: (*i*) the desire for the affect of the user to be positive, and (*ii*) the desire for the user to comply with the robot's suggestions. Robot drives are outcomes directly related to the robot that it wishes to incur including its own emotions and the progression of a bi-directional interaction: (*i*) a drive to complete the activity interaction, and (*ii*) a drive to be in a positive emotional state.

Emotional appraisal is achieved using the Ortony, Clore, and Collins (OCC) model, which describes how emotions arise as positive or negative responses to events, actions and objects [58]. The appraisal influence used by desires and drives is determined via the emotion generator defined in [82].

Observations of the robot's previous emotional expressions are used to determine whether a chosen deliberative emotional expression is displayed by the robot as expected or whether a reactive emotion is displayed instead. These observations influence the choice of future deliberative emotions.

The appraisal process is used to generate an appraised emotional vector $U_E = [ae_1 \dots ae_N]^T$, which represents the impact of the user's affect and their interaction speech; and an observation feedback vector $U_O = [oe_1 \dots oe_N]^T$, which represents the influence of the robot's own observed expressions. Together, these vectors update the emotional transition matrix, A_t. An initial transition matrix A_0 is updated by the new emotion transition matrix $A_t{}'$ as:

$$A_t' = \left(A_0^T \, U_E \, U_O\right)^T, \quad (8)$$

where ae_i represents the appraisal influence, and oe_i represents the observation feedback influence on emotion i, respectively. $A_t{}'$ is normalized to produce the emotional transition matrix A_t as follows:

$$A_t = A_t' \left[\frac{1}{\sum_i a_t^{i,1\prime}} \quad \frac{1}{\sum_i a_t^{i,2\prime}} \quad \dots \quad \frac{1}{\sum_i a_t^{i,N\prime}}\right]^T. \quad (9)$$

Appraisal influence: Appraisal influence ae_i is determined using the emotion generators in [82]. It focuses on determining the intensities of emotions associated with the robot's desires and drives. Desires and drives consist of a status, a level of priority and a likelihood of success, which are combined to determine these intensities. The status can be either *active*, *succeeded* or *failed*. Priority defines the level of importance of a desire or drive relative to others. Likelihood of success represents how likely an event or outcome is to occur based on previous knowledge about the event and the current sensory information. The appraisal influence for an emotion is described as:

$$ae_i = \sum_{D_S,} f_i(s_S^l, p_S^l, \mathcal{L}_S{}^l) + \sum_{D_R} g_i(s_R^h, p_R^h, \mathcal{L}_R{}^h), \quad (10)$$

where D_S is the set of desires $D_S = [ds_1 \dots ds_K]$, and D_R is the set of drives $D_R = [dr_1 \dots dr_G]$. f_i and g_i are the emotion generator functions for desires and drives, respectively. s_S^l is the status in the set $S = [Active, Succeeded, Failed]$ of desire l, and s_R^h is the status in the set S of drive h. p_S^l and $\mathcal{L}_S{}^l$ are the priority and likelihood of the desire l, respectively, with $l = 1 \dots K$. p_R^h and $\mathcal{L}_R{}^h$ are the priority and likelihood of the drive h respectively, with $h = 1 \dots G$.

Observation-feedback influence: The concurrence between deliberative and reactive emotions produces unpredictable robot emotion expression, hence, an observation module is required to verify if a chosen deliberative emotional expression is displayed by the robot as expected. Observation feedback influence, oe_i, is linked to these observations. If an expression is observed not to have been implemented as expected, the feedback biases the future emotional state towards repeating the emotion. oe_i for emotion i, is determined as:

$$oe_i = \begin{cases} w, & \text{where } w > 1 \\ 1 \end{cases}, \quad (11)$$

where it is equal to 1 (i.e., no influence) when the observed expression of the robot is the expected expression; and, it is equal to the influence weighting, w, when an observed expression is not implemented by the robot as expected. All other feedback influences are set to 1 in both cases.

3) Robot Reactive Emotions and Emotion Priority

The reactive emotions are determined as direct responses to unexpected situations and robot safety concerns during HRI. Namely, they allow the robot to respond to stimulus associated with potentially dangerous circumstances during HRI (e.g., injury, falling).

A rule-based reasoning approach is used to activate reactive emotions and their corresponding expressions:

$$\text{if } (S_i) \text{ then: } X_{t+1} = r_i, Y_{t+1} = rd_c, \quad (12)$$

where S_i represents the stimulus that activates reactive emotion i, r_i represents the reactive emotion i in the set of reactive emotions $R = [r_1 \dots r_Q]$ and rd_c represents the reactive expression i in the set of reactive expressions $RD = [rd_1 \dots rd_C]$.

A priority module in the REM is used to determine the final emotional expression that will be displayed based on the prevalence of robot safety concerns. Namely, priority is given to reactive emotions if the robot is in situations in which it may be harmed. This final emotion expression is, then, sent to the low-level controllers to be implemented with its bi-directional interaction behavior.

IV. An Interaction Activity

The proposed multimodal emotional HRI architecture can be applied to a number of different bi-directional HRI scenarios. One such HRI is presented in this paper. The objective of the IAS, considered herein, is to determine the appropriate robot

behavior to motivate a user to live a healthy lifestyle through meal and exercise planning. This is achieved by offering suggestions of meals and exercises each day. There exists a large body of evidence stating that healthy eating and regular exercises can help reduce the risk of chronic diseases in all stages of life [83]. However, it is not always easy for people to make this change on their own [84], therefore, a critical factor for change can be direct motivation [85]. We propose the use of a socially assistive robot to provide this motivation. In previous studies, the use of a person's affect to determine appropriate emotional behaviors for an assistant has resulted in the assistant being perceived as more empathetic and trustworthy [86].

A one-on-one multimodal HRI scenario is designed herein where a robot provides social assistance with this task. Autom is the only robot that has been designed to provide and monitor meal and exercise plans [52]. However, it utilizes user input provided through its tablet PC and does not engage in bi-directional affective communication.

In our experiments, interactions between the robot and users took place twice a day: once in the morning, where the robot makes recommendations for the rest of the day, and once in the evening, where the robot checks-in with the user. The behavior of the robot was designed using a finite-state machine for the morning and evening interactions. Examples of the robot's behaviors are shown in Table I.

Morning Interaction: At the start of the day, the robot greets the user and introduces itself, enquires about the weather and the user's dietary restrictions, and provides healthy-lifestyle meal and exercise suggestions.

Evening Interaction: At the end of the day, the robot carries out a social exchange, such as asking about the user's day and, then, enquires whether the suggested meals and exercise activities were complied with. The robot provides positive feedback if the user has eaten the suggested meals or healthy alternatives, and completed the suggested exercise plan and has been active. Otherwise, the robot encourages the user to follow the suggestions in the future.

V. The Social Robot, Luke

Our proposed HRI architecture was tested through a NAO robot platform, "Luke", developed by Aldebaran Robotics, Fig. 2. This robot has 25 degrees of freedom mobility, 8 RGB LEDs around each eye that are used to display multimodal emotional behavior, a synthesized voice that can be controlled via the pitch, speed and volume, touch sensors on its hands, feet and top of the head, and two cameras in its head.

A. Robot Emotions

The deliberative emotions we designed for our robot, Luke, included *happy*, *interested*, *sad*, *worried*, and *angry*. The reactive emotions were based on three types of *scared* that the robot would display in response to being picked up or touched by the user while providing assistance, or when it detected it was close to an edge that it could fall off. Due to the small form factor of the robot, Luke was placed on top of a table and the reactive emotions were used to ensure safe operation of the robot around an unpredictable user and table edges.

Each emotion had a set of expressions (high intensity and low intensity) defined by a unique combination of eye color, body language and vocal intonation. Body language was adapted from [87, 88], which also used a NAO robot. The eye color for each emotion was adapted from [89], and chosen to be unique for each emotion as in [89-91]. Example expressions are described in Table II and shown in Fig. 2.

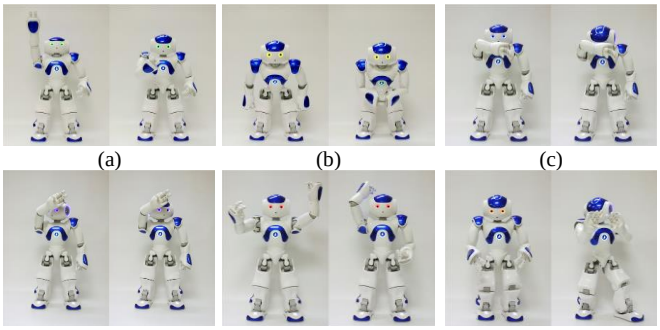

Fig 2. Luke's body-language expressions: (*a*) low-intensity happiness, (*b*) low-intensity interested, (*c*) high-intensity sadness, (*d*) high-intensity worried, (*e*) high-intensity anger, and (*f*) scared when touched.

It has been shown that vocal intonation [92], body language [93] and color [94] can provide distinct emotional information when used for basic emotions. We, thus, take advantage of these three modes and combine them in our robot emotional display to minimize ambiguity and the influence of such external factors such as ambient lighting and activation patterns [87, 89, 91]. Previous research has utilized the combination of body motion, color and sound together in robotic emotional display for both humanoid and non-humanoid robots [87, 90, 95]. User studies have shown that using such a combination of communication modes has improved emotion recognition and user confidence of an emotion when compared to unimodal emotion displays, as the combination minimizes classification errors and disambiguates resulting from single modes [90, 95]. For example, a specific eye color does not on its own represent a single emotion, since users can perceive different emotions from a given eye color [87, 91], with the exception being red for anger. Therefore, no single mode can effectively represent all

TABLE I
Example Robot Behaviors for the Diet and Fitness Activity

Behavior Type			Example
Social Exchange	Morning	Greet the user	"Hello Bob, how has your morning been so far?"
	Evening	Greet the user	"Hello again Bob, how has the rest of your day?"
Healthy-Life style Suggestions	Morning	Suggest a meal	"For breakfast, I suggest an apple and two slices of toast with strawberry jam."
		Suggest an exercise	"Since the weather is nice outside, you should take a walk; try walking five blocks one way, and back again."
		Determine dietary requirements	"Are you able to eat products containing gluten?"
Information Inquiries	Morning	Check-in regarding meals	"Did you eat the chicken on brown rice, with vegetables, for lunch today?"
	Evening	Check-in regarding exercise	"Did you walk the ten blocks outside today?"

emotions equally [95]. The utilization of the three modes of body motion, color and sound can increase recognition rates and user confidence in different robot emotions [90].

VI. System Training

Before conducting the experiments, we trained both the MARS and REM modules.

A. Training of the Multimodal Affect Recognition Classifier

A multimodal dataset with corresponding body language and vocal intonation samples was created for training the MARS using an approach similar to [96]. Namely, actors were asked to display eight emotional states (i.e., happy, sad, angry, fearful, surprise, disgust, calm, and neutral) while uttering two context-neutral statements ("kids are talking by the door" and "dogs are barking by the door"). We recruited seven student actors (five males and two females) to display these emotions while speaking the aforementioned statements. The body language classes used are adapted from [42, 65] and the vocal intonation rules for speech rate, prosody, fluency are from [97, 98]. Details of these body language and vocal intonation classes/rules are provided in Tables C and D in the supplementary material. Each actor recorded one session per emotion (total of eight sessions per actor). They performed the emotional display in front of the Kinect sensor and the microphone. After all recordings, each emotion display was segmented into 2-minute intervals to be coded by two expert coders. The expert coders rated both valence and arousal as (-2, -1, 0, 1, 2). Each segment in the database consisted of the coded valence and arousal values with the corresponding extracted features defined in Section III.A for body language and vocal intonation. In total, 929 samples were obtained. The class distribution for valence ranged from 15% (for both v=-1 and v=1) to 31% (for v=0), and for arousal ranged from 11% (for a=-1) to 32% (for a=0), with all other valence and arousal values within these ranges. A sample of the dataset is presented in Fig. A in the supplementary material.

The Bayesian network classifier, used within the multimodal affect recognition module, was trained using the database we created. We used ten-fold cross-validation, and achieved a classification rate of 86.8% for valence and 88.9% for arousal. The details of this validation test are presented in Section B of the supplementary material.

Our classification results are comparably higher than other multimodal emotion recognition systems. Namely, with respect to existing systems that use valence-arousal, their accuracy ranges from 52%-85% for valence and 71%-80% for arousal [99, 100]. For those using categorical emotions, the accuracy ranges from 43% to 83% [41, 100]. A detailed comparison table is provided in Table H in the supplementary material.

In order to further evaluate our emotion classification system, we trained our system using CreativeIT [101], a public dataset that provides multimodal (skeleton tracking and audio) human emotion displays coded using valence, arousal and dominance. Our system achieved an accuracy of 68.5% for valence and 70.4% for arousal, as well as a correlation of 0.528 for valence and 0.659 for arousal, which is compatible with results found in literature [102-104]. We provide more information in Section C of the supplementary material.

B. Training the REM

An initial learning stage using ten participants (ages 22-37, six males, four females) was performed to determine the initial values of A_0 and B. This learning stage was similar to the experiments, but with different participants. To minimize repetition during the compliance gaining behaviors, the robot would make a specific suggestion up to three times while rephrasing its speech. Initially, the deliberative emotions and their corresponding emotional expressions have the same probability of being chosen (uniform probability). Matrices A' and B' update the distribution probabilities as follows:

$$a'_{x,y} = a'_{x,y} + 1 \tag{13}$$

$$b'_{y,z} = b'_{y,z} + 1\,, \tag{14}$$

where the robot transitions from Emotion x to Emotion y, and then uses an emotional Expression z associated with y to provide a suggestion to the user. If the user accepts the suggestion, A' and B' are updated by (13) and (14). Otherwise, A' and B' are instead updated as follows:

$$a'_{x,i} = a'_{x,i} + \frac{1}{N-1}, i \in \{1 \dots N\}\backslash\{y\} \tag{15}$$

$$b'_{y,j} = b'_{y,j} + \frac{1}{w-1}, j \in \{1 \dots q\}\backslash\{z\}, \tag{16}$$

where N is the number of deliberative emotions and q is the size of the block for the emotion y. The matrix elements of A_0 and B are, then, updated by:

$$A_0 = \frac{A'}{C} \tag{17}$$

$$B = \frac{B'}{C}, \tag{18}$$

where C is the total number of suggestions given by the robot.

VII. Experiments

The proposed multimodal HRI architecture was evaluated through extensive experiments. The primary objective was to investigate the robot's ability to recognize user affect and adapt its own emotions based on the activity interaction.

A. Multimodal HRI Experiments

The goal of the social HRI experiments was to investigate the users' affect during an assistive interaction with Luke and evaluate how the users rated their experience with the robot. In

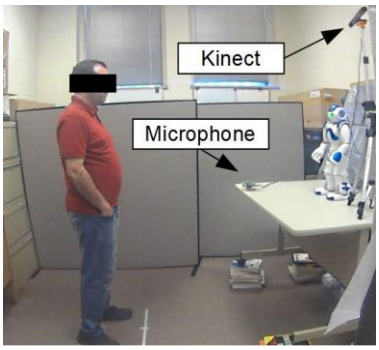

Fig. 3. Interaction Setup.

TABLE II
ROBOT MULTIMODAL AFFECTIVE DISPLAY

Affective State	Eye Color	Vocal Intonation	Body Language
Low-Intensity Happiness	Green	High pitch Fast tempo	The robot slowly looks upwards and raises its right arm and brings it down quickly (fist pump), Fig. 2a.
High-Intensity Happiness			The robot dances by swiftly swinging its hips from side to side, and moving its arms back and forth.
Low-Intensity Interested	Yellow	Fast tempo	The robot leans forward, looks up, and brings its hands together, Fig. 2b.
High-Intensity Interested			The robot leans forward, looks up, nods quickly and stretches both arms forwards.
Low-Intensity Sadness	Blue	Low pitch Slow tempo	The robot bends forward slowly and looks down.
High-Intensity Sadness			The robot slowly weeps into its forearm, Fig. 2c.
Low-Intensity Worried	Violet	High pitch	The robot crouches down, covers both of its eyes with its hands, and shakes its head slowly.
High-Intensity Worried			The robot puts the back of its hand on its forehead and shakes its head, Fig. 2d.
Low-Intensity Anger	Red	High pitch Fast tempo	The robot leans forward slightly with both hands on its hip and, then, shakes its head rapidly.
High-Intensity Anger			The robot waves its arms furiously in the air, Fig. 2e.
Scared, P, (Picked up)	Orange	High pitch Fast tempo	The robot freezes in place and remains stationary when being picked up by a user.
Scared, T, (Touched)			The robot quickly takes one step back leaning backwards and raises both hands to protect its face, Fig. 2f.
Scared, E (Edges)			The robot cover is eyes quickly with both hands.

order to investigate the influence of the robot's emotions on the interaction, we conducted a between-subjects experiment, with half of the users interacting with Luke using the emotion model (Emotionally Expressive Robot), and the other half with the robot not displaying any emotions (Neutral Robot). We also investigated which robot modality the users rated as more expressive after interacting with the robot.

The HRI experiments were conducted for the diet and fitness planning HRI application. In total, 18 university students between the ages of 20 to 42 participated in the experiments. The post-study questionnaire showed that participants were familiar with robots, but the majority had not interacted directly with a social robot before.

Luke was placed on top of a table at 0.75 m from the user for one-to-one standing interactions, Fig. 3. The microphone was placed in front of the robot on the table to capture a user's voice and the Kinect sensor was placed behind Luke to capture user body movements and poses during interaction. The touch sensors in the robot's arms are used to detect if the user is touching the robot, whereas the force sensitive resistors in the feet are used to detect if it was picked up. During the interaction, the embedded camera in the robot's mouth was used to track the user's face for feedback for robot gaze control (i.e., eye contact) during HRI.

During interactions with the Emotionally Expressive Luke and users, the robot autonomously detected a user's affect in real-time, determined its own emotional states and expressions, and implemented its appropriate emotional behavior based on the activity. The Neutral Luke used the same speech for the activity, however, without any body language or eye colors. Although the robot asked only closed-ended questions, the users might still give complex answers. Thus, an operator was utilized only for user speech recognition during the experiments to minimize reliability issues of current speech recognition and parsing software. The microphone was used to provide audio output to the operator who was located outside of the interaction environment and not visible to the participants.

At the end of each interaction, participants were asked to evaluate the application and their experience. The questionnaire was based on the Almere Model [54], and included seven questions on a 5-point Likert scale (1- strongly disagree to 5-strongly agree) to assess how users perceived the robot regarding (*i*) its ability to have sociable behavior, (*ii*) it being a social entity they are interacting with, (*iii*) its usefulness as a diet and fitness companion, and (*iv*) users trusting its advice. Finally, the questionnaire asked users to rank the robot's three emotional display modalities (eye color, body language, and vocal intonation) in terms of their ability to convey emotions effectively. This last question allows us to verify which modalities are considered crucial for an emotion-expressive socially assistive robot.

A video highlighting the interaction of our Emotionally Expressive Robot with different users is presented here[1] on our YouTube channel. When interacting with the user, the robot's emotions are displayed as a blended cue of body language, eye colors and emotionally intonated speech. Speech is generated using Nao's text-to-speech engine provided by Nuance.

B. *Experimental Results*

1) *Emotionally Expressive and Neutral Robot Comparison*

Valence and arousal were detected during the interactions in order to investigate whether Luke's emotional behavior had a positive influence on user affect. Fig. 4 shows the distribution of detected valence v and arousal a across participants during the interactions for both the Emotionally Expressive and Neutral robot. Both valence and arousal were recorded based on a scale of –2 (high negative) to +2 (high positive). In general, the participants had higher frequencies of neutral and positive valence (99%) and arousal (89%) than the Neutral robot (63% and 70% for valence and arousal, respectively). A Mann-Whitney U test showed that the differences between the valence and arousal distributions are statistically significant: Valence (U = 20505.5, p < 0.001, given α=0.001); and Arousal (U = 16598.5, p = 0.011, given α=0.05).

We performed an ensemble-average analysis, similar to [105], to investigate how the robot and user influenced each other's valence. We calculated the average valence displayed by the users and robot during each interaction stage in the morning and evening sessions. The robot emotions were converted to valence levels using the Circumplex Model [106].

On average, the users had higher valence when interacting

[1] https://youtu.be/COx1GxPV3_M

with the Emotionally Expressive Robot during both the morning and evening interactions (Fig. 5). For the majority of the interactions, their valence was positive while the robot's valence was also positive. When interacting with the Neutral Robot, on average, user valence was negative. The results show that the robot's positive emotions influenced the users' affect.

During the morning interactions, Fig. 5(a), the instances of higher average user valence for the Emotionally Expressive Robot were observed at the meal suggestion stage. Users displayed more positive valence when they were asked about their dietary requirements and when the robot was suggesting meals for lunch and dinner. Their body language was either neutral (i.e., with hands in pockets or down) or positive (nodding at the robot), and their vocal intonation had higher pitch, with some users laughing at the robot behavior.

During the evening interactions, Fig. 5(b), the decrease in user valence when interacting with the Emotionally Expressive Robot was a result of the users not eating dinner and not exercising. When giving the negative responses, users usually displayed neutral body language and vocal intonation.

The robot's positive emotions were determined using our proposed emotion model that directly considers user affect. As can be seen in Fig. 5, directly after the average user valence decreased, the emotionally expressive robot's valence increased in an attempt to improve user valence.

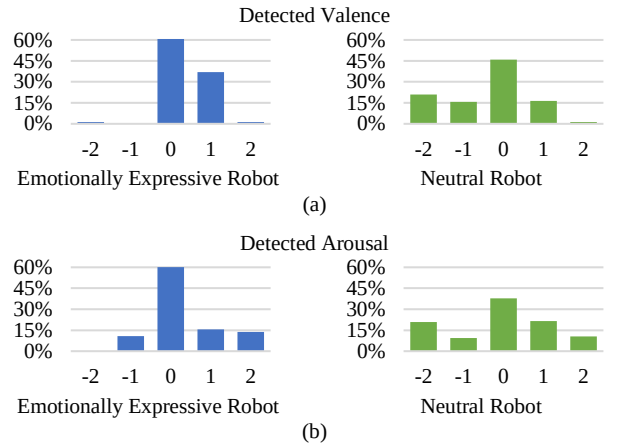

Fig. 4. Histogram of participants' (a) valence and (b) arousal, detected when interacting with the emotionally expressive robot and the neutral robot.

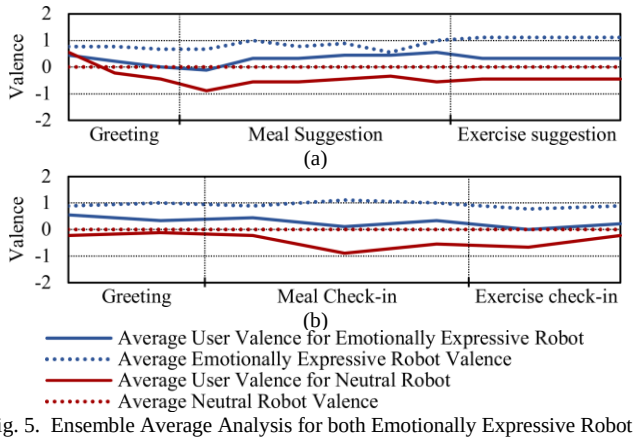

Fig. 5. Ensemble Average Analysis for both Emotionally Expressive Robot and Neutral Robot during (a) morning and (b) evening interactions.

2) Questionnaire Results

The questionnaire results for the overall interaction experience with the Emotionally Expressive and Neutral robots are in Table III. In general, the participants who interacted with the emotional robot found the interaction more pleasant and the emotions real and understandable. Based on their responses, it was clear that they were aware of the robot's emotions.

The questionnaire also evaluated how users perceived the usefulness of the Emotionally Expressive and Neutral robots as a diet and fitness companion. The results indicate that both the Emotionally Expressive and Neutral Robots achieved similar usefulness rates. This is expected, as both robots used the same interaction stages and speech. In general, the users were neutral after the use of either robot for the day with respect to their intent to use the robot again and listen to its advice. They mentioned that they wanted the robot to provide more individualized exercise suggestions, with the option to follow a long-term plan to improve their overall lifestyle.

The ranking for the emotional display modes that the Emotionally Expressive robot used is presented in Table IV, where "1" is the highest rank. Participants were allowed to use the same rank for multiple modes. Vocal intonation was the dominant modality for displaying the robot's affect (77.8% participant agreement), followed by body language (33.3% participant agreement), and eye colors (33.3% participant agreement). As shown in Table IV, body language had a slightly lower (three units) rank sum than eye color, which is why the participant agreement for both was the same.

3) Accuracy of Affect Detection

To evaluate the affect recognition rate, two expert coders (one male, one female) independently coded the valence and arousal levels of all the participants using body language and vocal intonation information. Both coders were presented with videos of the interactions for independent coding. The coders, then, met to discuss their results to obtain inter-coder consensus in order to reduce coder bias [107]. The videos were synchronized with the inputs used for affect classification by

TABLE III
ROBOT PERFORMANCE QUESTIONS

Question	Neutral		Emotional	
	Mean	SD	Mean	SD
I feel that I understood the emotions that Luke displayed	2.8	1.2	4.1	0.3
Sometimes Luke seems to have real feelings	2.3	1.0	3.4	1.0
I find Luke pleasant to interact with	3.4	1.3	4.0	0.7
I think Luke is nice	4.0	0.9	4.6	0.5
I feel like Luke understands me	3.0	1.2	2.9	1.0
I would follow the advice Luke gives me	3.0	1.2	3.0	1.2
I would use Luke as a diet and fitness companion	3.0	1.3	3.0	1.2

SD: Standard Deviation

TABLE IV
RANKING OF EMOTIONAL DISPLAYS

Robot Modality	Rank	Rank Sum	Participant agreement
Vocal Intonation	1	11	77.8%
Body Language	2	18	33.3%
Eye Color	3	21	33.3%

the robot that were performed during the experiments to obtain the corresponding valence and arousal values.

The accuracy rate of the system was determined to be 76.0% for valence and 60.4% for arousal, respectively. Our system had lower performance with the experimental data due to the inaccuracy of the OpenNI/NITE in detecting the skeleton joints, since it could not properly detect small movements (e.g., "feet stamping" or "subtle nodding") or poses with occlusion (e.g., "arms crossed in front of the torso", "legs crossed" or "arms behind the trunk"). This, particularly, affected the arousal classification, which required the detection of subtle movements to detect positive arousal.

The REM coped with inaccuracies due to affect detection by also considering additional knowledge about the interaction (i.e., user compliance, and completing the activity interaction) and its own previous emotional states through its other desires and drives. It did not solely rely on the desire for the user to be positive when determining its current emotional state.

4) Robot Emotions

For the Emotionally Expressive Robot, we also investigated the relationship between the user affect and the robot emotional display. Fig. 6 shows the affect levels of the participants and the robot's corresponding emotions during both morning and evening interactions. In particular, in Fig. 6, we highlighted the interactions of Users 1 to 5 in different colors, with the remaining users' interactions in gray.

Regarding the general transitions of robot emotional states during the morning interaction, Fig. 6(a), the majority of changes occurred while greeting users and suggesting meals and exercises. These changes were in accordance with the robot's desire to (*i*) improve the user affect and (*ii*) to obtain user compliance. For example, the robot transitioned from the interested state to the happy state when User 4 accepted the suggested lunch. Since most users agreed to the robot's suggestions, it mostly displayed happy and interested emotions. The robot transitioned from interested to the sad state when User 2 did not agree with a meal suggestion, but transitioned back to interested once User 2 agreed to eat an alternative meal. The robot displayed one of the reactive scared states to User 2 when its onboard camera detected that it was close to the edge of the table. The robot asked that the user assist it by moving it back away from the edge. Once the robot identified it was away from the edge, it was no longer scared and transitioned back to the interested state.

Regarding the transitions of robot emotional states during the afternoon interaction, Fig. 6(b), most changes in the robot's states occurred while the robot was checking if the users ate the suggested meals or completed the suggested exercises. This is consistent with the robot's desire to achieve compliance with its suggestions. The robot entered the happy or the interested state when users complied (i.e., the desire for user compliance succeeded), and the sad or worried state when users did not comply (i.e., the desire failed). To highlight the influence user compliance had on the robot's emotional states, it can be noted that the robot became worried after User 2 said that he/she skipped lunch. For Users 3 and 5, the robot transitioned to a happy state when they did the suggested exercise.

In order to highlight the robot behavior, when interacting with a specific user, we analyze herein the interaction with User 1. During the morning interaction, the robot greeted User 1 in a low-intensity happy state, then, transitioned to a high-intensity interested state when the user responded that the weather was nice outside. The robot became sad when the user stated that he/she did not have breakfast and, then, suggested to the user a breakfast option. The robot was in an interested state for the lunch and dinner suggestions, and transitioned to a happy state for the exercise suggestion, since the user agreed to all of the robot's suggestions. Regarding user affect during the first half of the morning interaction, the user was in a more neutral state and appeared to be focused on the information being provided by the robot. However, the user showed more high-energy body gestures and voice (positive arousal), and open and stretched body movements and high level of content in the voice (positive valence) during the second half of the interaction.

During the evening interaction, for the same User 1, the robot greeted him/her in a low-intensity happy state and transitioned to the high-intensity interested state when the user responded that the day was going great. Then, the robot transitioned to a low-intensity sad state when it found out the user did not eat the suggested breakfast, and became worried when user responded that he/she decided to not have breakfast. However, the robot transitioned back to an interested state when the user responded that he/she ate the suggested lunch. When checking-in about dinner, the robot was in a high-intensity happy state by displaying its dance. By doing so it approached the edge of the table, activating the reactive scared emotion. After the user

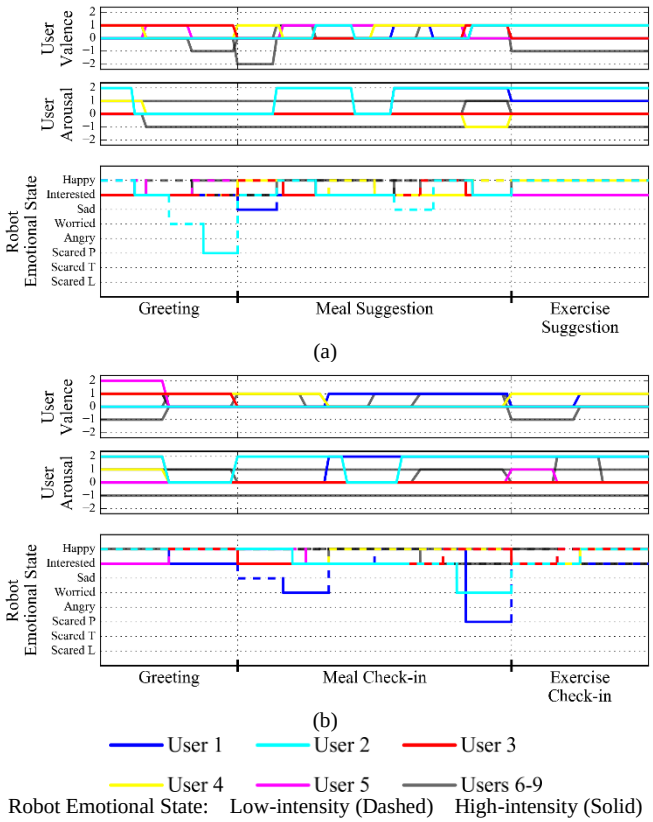

Fig. 6. User affect and robot emotional expressions during (a) morning, and (b) evening interactions for all participants.

helped move Luke to the center of the table, the robot transitioned to the interested emotional state and performed the exercise check-in. With respect to user affect, similar to the morning interaction, the user displayed neutral valence and arousal during the first half of the interaction. During the second half of the interaction, the user displayed positive valence and high positive arousal when discussing his/her lunch and dinner meals. Positive arousal and valence were also detected during the exercise checking-in stage when the user laughed at the robot when it told an exercise joke.

VIII. Conclusions

In this paper, a novel multimodal emotional HRI architecture is presented for effective bi-directional communication between a user and a robot. It allows a robot to both determine a user's affect using the unique combination of body language and vocal intonation and, in turn, determine its own appropriate emotional behavior using a two-layer robot emotional model.

We verified our architecture with a small humanoid robot to investigate its ability to detect affect, and adapt its emotion to changes in user affect and the progression of the interaction at hand during a diet and fitness counselling HRI scenario. Experimental results clearly verified that the Emotionally Expressive Robot can induce more positive valence and less negative arousal in users when compared to the Neutral Robot. Questionnaires also indicated that users considered the Emotionally Expressive Robot to be more enjoyable to interact with, which was also evident by their more positive valence.

As future work, we will study how a social robot can influence people's diet and exercise routines during long-term interactions and how this can impact their overall lifestyle.

References

[1] M. Gazzaniga, T. Heatherton, and D. Halpern, "Motivation and Emotion," *Psychol. Science*, 4th ed., New York, NY, USA: W. W. Norton & Company, pp. 403-450, 2012.

[2] D. McColl, A. Hong, N. Hatakeyama, G. Nejat, and B. Benhabib, "A Survey of Autonomous Human Affect Detection Methods for Social Robots Engaged Natural HRI," *J. Intell. Robot. Syst. Theory Appl.*, vol. 82, no. 1, pp. 101-133, Apr. 2016.

[3] J. M. K. Westlund, H. W. Park, R. Williams, and C. Breazeal, "Measuring young children's long-term relationships with social robots," *17th ACM Conf. Interact. Des. Children*, pp. 207-218, 2018.

[4] N. Mavridis *et al*., "FaceBots: Steps Towards Enhanced Long-Term Human-Robot Interaction by Utilizing and Publishing Online Social Information," *Paladyn*, vol. 1, no. 3, pp. 169-178, 2010.

[5] N. Mitsunaga, T. Miyashita, H. Ishiguro, K. Kogure and N. Hagita, "Robovie-IV: A Communication Robot Interacting with People Daily in an Office," *IEEE Int. Conf. Intel. Robots Syst.*, pp. 5066-5072, 2006.

[6] L. D. Riek *et al.*, "Ibn sina steps out: Exploring arabic attitudes toward humanoid robots," *Int. Symp. New Frontiers Human–Robot Interact.*, pp. 88-94, 2010.

[7] C. Becker-Asano and H. Ishiguro, "Evaluating facial displays of emotion for the android robot Geminoid F," *IEEE Workshop Affect. Comput. Intell.*, pp. 1-8, 2011.

[8] J. Li, S. Qiu, Y. Shen, C. Liu and H. He, "Multisource Transfer Learning for Cross-Subject EEG Emotion Recognition," *IEEE Trans. Cybern.*, [Early access].

[9] M. J. Han, C. H. Lin, and K. T. Song, "Robotic emotional expression generation based on mood transition and personality model," *IEEE Trans. Cybern.*, vol. 43, no. 4, pp. 1290-1303. Aug. 2013.

[10] L. Xin, X. Lun, W. Zhi-Liang, and F. Dong-Mei, "Robot emotion and performance regulation based on HMM," *Int. J. Adv. Robot. Syst.*, vol. 10, no. 3, p.160, Mar. 2013.

[11] X. Zhao, J. Zou, H. Li, E. Dellandrea, I. A. Kakadiaris, and L. Chen, "Automatic 2.5-D Facial Landmarking and Emotion Annotation for Social Interaction Assistance," *IEEE Trans. Cybern.*, vol. 46, no. 9, pp. 2042-2055, Sep. 2016.

[12] K. Mistry, L. Zhang, S. C. Neoh, C. P. Lim, and B. Fielding, "A Micro-GA Embedded PSO Feature Selection Approach to Intelligent Facial Emotion Recognition," *IEEE Trans. Cybern.*, vol. 47, no. 6, pp.1496-1509, Jun. 2017.

[13] L. Zhong, Q. Liu, P. Yang, J. Huang, and D. N. Metaxas, "Learning Multiscale Active Facial Patches for Expression Analysis," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1499-1510, Aug. 2015.

[14] L. Chen, M. Wu, M. Zhou, Z. Liu, J. She, and K. Hirota., "Dynamic Emotion Understanding Human-Robot Interaction Based on Two-Layer Fuzzy SVR-TS Model," *IEEE Trans. Syst., Man, Cybern., Syst*., vol. 50, no. 2, pp. 490-501, Feb. 2020.

[15] L. Chen, M. Zhou, W. Su, M. Wu, J. She, and K. Hirota, "Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction," *Info. Sci.*, vol. 428, pp. 49–61, 2018.

[16] L. Chen *et al.* "Three-Layer Weighted Fuzzy Support Vector Regression for Emotional Intention Understanding in Human–Robot Interaction," *IEEE Tran. Fuzzy Syst.*, vol. 26, no. 5, pp. 2524–2538, 2018.

[17] L. Chen, M. Wu, M. Zhou, J. She, F. Dong and K. Hirota, "Information-Driven Multirobot Behavior Adaptation to Emotional Intention in Human–Robot Interaction," *IEEE Trans. Cognit. Dev. Syst.*, vol. 10, no. 3, pp. 647-658, Sept. 2018.

[18] Y. Zong, W. Zheng, Z. Cui, G. Zhao and B. Hu, "Toward Bridging Microexpressions From Different Domains," *IEEE Trans. Cybern.*, [Early access].

[19] S. Deb and S. Dandapat, "Multiscale Amplitude Feature and Significance of Enhanced Vocal Tract Information for Emotion Classification," *IEEE Tran. Cyber.*, vol. 49, no. 3, pp. 802-815, March 2019.

[20] L. Chen, W. Su, Y. Feng, M. Wu, J. She, and K. Hirota, "Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction," *Info. Sci*., vol. 509, pp. 150-163, Jan. 2020.

[21] A. Hong, Y. Tsuboi, G. Nejat, and B. Benhabib, "Affective Voice Recognition of Older Adults," *ASME J. Med. Dev.*, vol. 10, no. 2, pp. 020931-1–020931-2, 2016.

[22] D. McColl, Z. Zhang, and G. Nejat, "Human body pose interpretation and classification for social human-robot interaction," *Int. J. Soc. Robot.*, vol. 3, no, 3, p. 313, Aug. 2011.

[23] C. Breazeal, "Social interactions in HRI: The robot view," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 34, no. 2, pp. 181-186, 2004.

[24] M. Ficocelli, J. Terao, and G. Nejat, "Promoting Interactions between Humans and Robots Using Robotic Emotional Behavior," *IEEE Trans. Cybern.*, vol. 46, no. 12, pp. 2911-2923, Dec. 2016.

[25] D. McColl, C. Jiang, and G. Nejat, "Classifying a Person's Degree of Accessibility From Natural Body Language During Social Human-Robot Interactions," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 524-538, 2017.

[26] D. McColl and G. Nejat, "Affect detection from body language during social HRI," *IEEE Int. Workshop Robot Human Interact. Commun.*, pp. 1013-1018, Sep. 2012.

[27] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 10, pp.1175-1191, Oct. 2001.

[28] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18-37, Jan. 2010.

[29] M. Paleari, R. Chellali, and B. Huet, "Bimodal Emotion Recognition," *Int. Conf. on Soc. Robotics*, pp. 305-314, Nov. 2010.

[30] C. Breazeal and R. Brooks, "Robot emotions: A Functional perspective," *Who Needs Emotion?: Brain Meets Robot*, pp. 271-310, 2005.

[31] D. K. Limbu *et al*., "Affective social interaction with CuDDler robot," *IEEE Conf. Robot., Automat. Mechatronics*, 2013, pp. 179-184.

[32] C. P. Lee-Johnson and D. A. Carnegie, "Mobile robot navigation modulated by artificial emotions," *IEEE Trans. Syst. Man, Cybern. Part B Cybern.*, vol. 40, no. 2, pp. 469-480, Apr. 2010.

[33] J. Saldien, K. Goris, B. Vanderborght, J. Vanderfaeillie, and D. Lefeber, "Expressing emotions with the social robot probo," *Int. J. Soc. Robot.*, vol. 2, no. 4, pp. 377-389, Dec. 2010.

[34] C. Breazeal, "Emotive qualities in robot speech," *IEEE/RSJ Int. Conf. on Intell. Robots and Syst.*. pp. 1388-1394, 2001.

[35] D. McColl and G. Nejat, "Recognizing Emotional Body Language Displayed by a Human-like Social Robot," *Int. J. Soc. Robot.*, vol. 6, no. 2, pp. 261-280, Apr. 2014.

[36] F. Cid, J. Moreno, P. Bustos, and P. Núñez, "Muecas: A multi-sensor robotic head for affective human robot interaction and imitation," *Sensors*, vol. 14, no. 5, pp. 7711-7737, Apr. 2014.

[37] H.-W. Jung, Y.-H. Seo, M. S. Ryoo, and H. S. Yang, "Affective communication system with multimodality for a humanoid robot, AMI," *IEEE/RAS Int. Conf. Humanoid Robot.*, pp. 690–706, Nov. 2004.

[38] J. A. Prado, C. Simplício, N. F. Lori, and J. Dias, "Visuo-auditory Multimodal Emotional Structure to Improve Human-Robot-Interaction," *Int. J. Soc. Robot.*, vol. 4, no. 1, pp. 29-51, Jan. 2012.

[39] T. Belpaeme *et al.*, "Multimodal Child-Robot Interaction: Building Social Bonds," *J. Human-Robot Interact.*, vol. 1, no. 2, pp. 33-53, Jan. 2013.

[40] D. McColl, W. G. Louie and G. Nejat, "Brian 2.1: A socially assistive robot for the elderly and cognitively impaired," *IEEE Robot. Automat. Mag.*, vol. 20, no. 1, pp. 74-83, March 2013.

[41] F. Alonso-Martín, M. Malfaz, J. Sequeira, J. F. Gorostiza, and M. Á. Salichs, "A multimodal emotion detection system during human-robot interaction," *Sensors*, vol. 13, no. 11, pp. 15549-15581, Nov. 2013.

[42] H. G. Wallbott, "Bodily expression of emotion," *Eur. J. Soc. Psychol.*, vol. 28, no. 6, pp. 879-896, Nov. 1998.

[43] H. Meng and N. Bianchi-Berthouze, "Affective state level recognition in naturalistic facial and vocal expressions," *IEEE Trans. Cybern.*, vol. 44, no. 3, pp. 315-328, Mar. 2014.

[44] N. Lazzeri, D. Mazzei, and D. De Rossi, "Development and Testing of a Multimodal Acquisition Platform for Human-Robot Interaction Affective Studies," *J. Human-Robot Interact.*, vol. 3, no. 2, pp. 1-24, Jul. 2014.

[45] A. Lim and H. G. Okuno, "The MEI robot: Towards using motherese to develop multimodal emotional intelligence," *IEEE Trans. Auton. Mental Develop.*, vol. 6, no. 2, pp. 126-138, Jun. 2014.

[46] H. Yang, Z. Pan, M. Zhang, and C. Ju, "Modeling emotional action for social characters," *Knowl. Eng. Rev.*, vol. 23, no. 4, pp. 321-337, Dec. 2008.

[47] R. Kirby, J. Forlizzi, and R. Simmons, "Affective social robots," *Rob. Auton. Syst.*, vol. 58, no. 3, pp. 322-332, Mar. 2010.

[48] X. Hu, L. Xie, X. Liu, and Z. Wang, "Emotion expression of robot with personality," *Math. Problems Eng.*, vol. 2013, pp. 1-10, 2013.

[49] J. C. Park, H. R. Kim, Y. M. Kim, and D. S. Kwon, "Robot's individual emotion generation model and action coloring according to the robot's personality," *IEEE Int. Workshop Robot Human Interact. Commun.*, pp. 257-262, Sep. 2009.

[50] X. Zhang, S. F. R. Alves, G. Nejat, B. Benhabib, "A Robot Emotion Model with History," *IEEE Int. Symp. Robot. Intell. Sensors*, pp. 230-235, Oct. 2017.

[51] K. Terada and C. Takeuchi, "Emotional Expression in Simple Line Drawings of a Robots Face Leads to Higher Offers in the Ultimatum Game," *Frontiers Psychol.*, vol. 8, pp. 1-9, 2017.

[52] C. D. Kidd and C. Breazeal, "Robots at home: Understanding long-term human-robot interaction," *IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pp. 3230-3235, Sep. 2008.

[53] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. W. Mcowan, "Multimodal affect modeling and recognition for empathic robot companions," *Int. J. Humanoid Robot.*, vol. 10, no. 1, pp. 1350010-1–1350010-23, Mar. 2013.

[54] M. Heerink, B. Kröse, V. Evers, and B. Wielinga, "Assessing acceptance of assistive social agent technology by older adults: The Almere model," *Int. J. Soc. Robot.*, vol. 2, no. 4, pp. 361-375, Dec. 2010.

[55] J. C. Castillo, Á. Castro-González, F. Alonso-Martín, A. Fernández-Caballero, M. Á. Salichs, "Emotion detection and regulation from personal assistant robot in smart environment," *Pers. Assistants: Emerg. Comput. Technol.*, A. Costa, V. Julian and P. Novais Eds. Cham, Switzerland: Springer, 2018, pp. 179-195.

[56] L. Paletta *et al.*, "AMIGO - towards social robot based motivation for playful multimodal intervention in dementia," *Pervasive Technol. Related Assistive Environ. Conf.*, pp. 421–427, Jun. 2018.

[57] J. Chan and G. Nejat, "Social Intelligence for a Robot Engaging People in Cognitive Training Activities," *Int. J. Advanced Robot. Sys.: Human-Robot Interact.*, vol. 9, pp. 1-12, 113:2012.

[58] B. N. Colby, A. Ortony, G. L. Clore, and A. Collins, "The Cognitive Structure of Emotions," *Contemp. Sociol.*, May 1989.

[59] J. A. Russell, A. Weiss, and G. A. Mendelsohn, "Affect Grid: A Single-Item Scale of Pleasure and Arousal," *J. Pers. Soc. Psychol.*, vol. 57, no. 3, pp. 493-502, Sep. 1989.

[60] A. K. Pérez, C. A. Quintero, S. Rodríguez, E. Rojas, O. Peña, and F. Rosa, "Identification of Multimodal Signals for Emotion Recognition in the Context of Human-Robot Interaction," *Int. Symp. Intell. Comput. Syst.*, pp. 67-80, 2018.

[61] L. A. Perez-Gaspar, S. O. Caballero-Morales, and F. Trujillo-Romero, "Multimodal emotion recognition with evolutionary computation for human-robot interaction," *Expert Syst. Appl.*, vol. 66, pp. 42-61, 2016.

[62] M. Shao, S. F. R. Alves, O. Ismail, X. Zhang, G. Nejat and B. Benhabib, "You Are Doing Great! Only One Rep Left: An Affect-Aware Social Robot for Exercising," *IEEE Int. Conf. Sys. Man Cybern*., pp. 3811-3817, 2019.

[63] Z. T. Liu *et al.*, "A multimodal emotional communication based humans-robots interaction system," *Chinese Control Conf.*, pp. 6363-6368, Jul. 2016.

[64] C. Shan, S. Gong, and P. W. McOwan, "Beyond Facial Expressions: Learning Human Emotion from Body Gestures," *British Mach. Vision Conf.*, pp. 1-10, Sep. 2007.

[65] M. de Meijer, "The contribution of general features of body movement to the attribution of emotions," *J. Nonverbal Behav.*, vol. 13, no. 4, pp. 247-268, Dec. 1989.

[66] M. Coulson, "Attributing emotion to static body postures: Recognition accuracy confusions and viewpoint dependence", *J. Nonverbal Behav.*, vol. 28, no. 2, pp. 117-139, 2004.

[67] N. Bianchi-Berthouze, P. Cairns, A. L. Cox, "On posture as a modality for expressing and recognizing emotions", *Emotion HCI Workshop*, pp. 74-80, 2006.

[68] A. P. Atkinson, W. H. Dittrich, A. J. Gemmell, and A. W. Young, "Emotion Perception from Dynamic and Static Body Expressions in Point-Light and Full-Light Displays," *Perception*, vol. 33, no. 6, pp. 717–746, 2004.

[69] K. R. Scherer, "Vocal Affect Expression. A Review and a Model for Future Research," *Psychol. Bull.*, vol. 99, no. 2, pp. 143-165, 1986.

[70] P. Ekman, "Darwin, Deception, and Facial Expression," *Ann. New York Acad. Sci.*, vol. 1000, no. 1, pp. 205–221, 2006.

[71] M. Okubo, A. Kobayashi, and K. Ishikawa, "A fake smile thwarts cheater detection," *J. Nonverbal Behav.*, vol. 36, no. 3, pp. 217-225, 2012.

[72] H. Aviezer, Y. Trope and A. Todorov, "Body cues, not facial expressions, discriminate between intense positive and negative emotions," *Science*, vol. 338, no. 6111, pp. 1225-1229, Nov. 2012.

[73] M. Zuckerman, D. T. Larrance, N. H. Spiegel, and R. Klorman, "Controlling nonverbal displays: Facial expressions and tone of voice," *J. Exp. Soc. Psychol.*, vol. 17, no. 5, pp. 506–524, 1981.

[74] L. Anolli and R. Ciceri, "The voice of deception: Vocal strategies of naive and able liars," *J. of Nonverbal Behav.*, vol. 21, no. 4, pp. 259–284, 1997.

[75] L. F. Barrett, "Discrete Emotions or Dimensions? The Role of Valence Focus and Arousal Focus," *Cogn. Emotion*, vol. 12, no. 4, pp. 579-599, 1998.

[76] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Develop. Psychopathol.*, vol. 17, no. 3, pp. 715-734, Sep. 2005.

[77] D. McColl and G. Nejat, "Determining the affective body language of older adults during socially assistive HRI," *IEEE Int. Conf. Intell. Robots Syst.*, pp. 2633-2638, Sep. 2014.

[78] D. McColl and G. Nejat, "A Socially Assistive Robot that can Monitor Affect of the Elderly during Meal-Time Assistance," *ASME J. Med. Dev.*, vol. 8, no. 3, 030941, 2014.

[79] J. Shotton *et al.*, "Real-time human pose recognition in parts from single depth images," *Mach. Learn. Comput. Vision*, R. Cipolla, S. Battiato, G. Farinella Eds. Berlin: Springer, 2013, pp. 116-124.

[80] A. Liberman, "Apparatus and methods for detecting emotions," *U.S. Patent No. 6,638,217*, issued October 28, 2003.

[81] Nemesysco, *QA5 SDK product description and user guide*, Nemesysco, Netanya, Israel, 2009.

[82] W. S. N. Reilly, "Believable Social and Emotional Agents," PhD Thesis, Carnegie-Mellon Univ., Pittsburgh, PA, 1996.

[83] U.S. Department of Health and Human Services and U.S. Department of Agriculture. *2015–2020 Dietary Guidelines for Americans*, 8th ed., Dec. 2015. [Online]. Available: http://health.gov/dietaryguidelines/2015/guidelines/ [Accessed: Oct. 10, 2018].

[84] J. Sargeant, M. Valli, S. Ferrier, and H. MacLeod, "Lifestyle counseling in primary care: Opportunities and challenges for changing practice," *Med. Teacher*, vol. 30, no. 2, pp. 185-191, Jan. 2008.

[85] W. R. Miller and S. Rollnick, *Motivational interviewing: preparing people for change*, New York, NY: The Guilford Press, 2002, pp. 201-206.

[86] R. Looije, M. A. Neerincx, and F. Cnossen, "Persuasive robotic assistant for health self-management of older adults: Design and evaluation of social behaviors," *Int. J. Human Comput. Stud.*, vol. 68, no. 6, pp.

386-397, 2010.
[87] M. Haring, N. Bee, and E. Andre, "Creation and Evaluation of emotion expression with body movement, sound and eye color for humanoid robots," *IEEE Int. Workshop Robot Human Interac. Commun.*, pp. 204-209, Jul. 2011.
[88] A. Beck, L. Cañamero and K. A. Bard, "Towards an Affect Space for robots to display emotional body language," *19th IEEE Int. Symp. Robot Human Interact. Commun.*, Viareggio, pp. 464-469, 2010.
[89] K. Terada, A. Yamauchi and A. Ito, "Artificial emotion expression for a robot by dynamic color change," *21st IEEE Int. Symp. Robot Human Interact. Commun.*, Paris, pp. 314-321, 2012.
[90] S. Song, and S. Yamada, "Expressing emotions through color, sound, and vibration with an appearance-constrained social robot," *ACM/IEEE Int. Conf. Human-Robot Interact.*, pp. 2-11, 2017.
[91] D. O. Johnson, R. H. Cuijpers, and D. V. D. Pol, "Imitating Human Emotions with Artificial Facial Expressions," *Int. J. Soc. Robot.*, vol. 5, no. 4, pp. 503–513, 2013.
[92] K. R. Scherer, R. Banse, and H. G. Wallbott, "Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures," *J. Cross-Cultural Psychol.*, vol. 32, no. 1, pp. 76–92, 2001.
[93] Z. Witkower and J. L. Tracy, "Bodily Communication of Emotion: Evidence for Extrafacial Behavioral Expressions and Available Coding Systems," *Emotion Rev.*, vol. 11, no. 2, pp. 184–193, 2018.
[94] X.-P. Gao, J. H. Xin, T. Sato, A. Hansuebsai, M. Scalzo, K. Kajiwara, S.-S. Guan, J. Valldeperas, M. J. Lis, and M. Billger, "Analysis of cross-cultural color emotion," *Color Res. Appl.*, vol. 32, no. 3, pp. 223–229, 2007.
[95] D. Löffler, N. Schmidt, and R. Tscharn, "Multimodal Expression of Artificial Emotion in Social Robots Using Color, Motion and Sound," *ACM/IEEE Int. Conf. Human-Robot Interact.*, pp. 334-343, 2018.
[96] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *Plos One*, vol. 13, no. 5, 2018.
[97] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression.," *J. Pers. Soc. Psychol.*, vol. 70, no. 3, pp. 614–636, 1996.
[98] D. A. Sauter, F. Eisner, A. J. Calder, and S. K. Scott, "Perceptual Cues in Nonverbal Vocal Expressions of Emotion," *Quart. J. Exp. Psychol.*, vol. 63, no. 11, pp. 2251–2272, 2010.
[99] W. Liu, W. Zheng, and B. Lu. "Emotion recognition using multimodal deep learning," *Int. Conf. Neural Info. Process.*, pp. 521-529, 2016.
[100] C.-C. Lu, J.-L. Li, and C.-C. Lee, "Learning an Arousal-Valence Speech Front-End Network using Media Data In-the-Wild for Emotion Recognition," *ACM Aud./Vis. Emotion Chall. Workshop*, pp. 99-105, 2018.
[101] A. Metallinou, Z. Yang, C. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The USC CreativeIT database of multimodal dyadic interactions: from speech and full body motion capture to continuous emotional annotations," *J. Lang. Resour. Eval.*, vol. 50, no. 3, pp. 497-521, 2016.
[102] C. Chang and C. Lee, "Fusion of multiple emotion perspectives: Improving affect recognition through integrating cross-lingual emotion information," *IEEE Int. Conf. Acoust., Speech Signal Process.*, pp. 5820-5824, 2017.
[103] A. Metallinou, A. Katsamanis, Y. Wang and S. Narayanan, "Tracking changes in continuous emotion states using body language and prosodic cues," *IEEE Int. Conf. Acoust., Speech Signal Process.*, pp. 2288-2291, 2011.
[104] A. Metallinou, A. Katsamanis, and S. Narayanan, "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information," *Image Vision Comput.*, vol. 31, no 2, pp. 137-152, 2013
[105] S. Costa, A. Brunete, B.-C. Bae, and N. Mavridis, "Emotional Storytelling Using Virtual and Robotic Agents," *Int. J. Humanoid Robot.*, vol. 15, no. 03, pp. 1850006-1–1850006-31, 2018.
[106] J. A. Russell, "A circumplex model of affect," *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161-1178, 1980.
[107] J. Trace, G. Janssen, and V. Meier, "Measuring the impact of rater negotiation in writing performance assessment," *Lang. Testing*, vol. 34, no. 1, pp. 3–22, 2016.

Alexander Hong received his B.A.Sc. degree in Engineering Science in 2014 and his M.A.Sc. degree in Mechanical Engineering in 2016, under the supervision of Prof. Nejat and Prof. Benhabib, from the University of Toronto (UofT), Toronto, Canada. His research interests include affect recognition in healthcare robots and semi-autonomous robot teams in urban search and rescue.

Nolan Lunscher was a research assistant working on robot emotion models.

Tianhao Hu was a research assistant working on robot emotion displays.

Yuma Tsuboi was an M.Eng student working on autonomous affect detection.

Xinyi Zhang received her M.A.Sc. degree in the Department of Mechanical & Industrial Engineering at UofT in 2019, under the supervision of Prof. Nejat and Prof. Benhabib. Her research interests include humanoid robotics, emotion calculation and autonomous systems.

Silas Franco dos Reis Alves received his Ph.D. degree in Electrical Engineering in 2016. He is currently a Post Doctoral Fellow in the Autonomous Systems and Biomechatronics Laboratory (ASBLab) at the University of Toronto. His research interests include socially assistive robots, assistive technologies, human-robot interactions and intelligent control architectures.

Goldie Nejat (S'03-M'06) is the Canada Research Chair in Robots for Society and a Professor in the Department of Mechanical & Industrial Engineering at UofT. She is the Founder and Director of the ASBLab. She is also an Adjunct Scientist at the Toronto Rehabilitation Institute. Her research interests include intelligent assistive/service robots, human-robot interactions, and semi-autonomous/ autonomous control. She received her B.A.Sc. and Ph.D. degrees in Mechanical Engineering at the University of Toronto.

Beno Benhabib received the B.Sc., M.Sc., and Ph.D. degrees in mechanical engineering in 1980, 1982, and 1985, respectively. He has been a Professor with the Department of Mechanical and Industrial Engineering at UofT since 1986. His current research interests include the design and control of intelligent autonomous systems.