This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Toward Artificial Emotional Intelligence for Cooperative Social Human–Machine Interaction

Berat A. Erol, *Student Member, IEEE*, Abhijit Majumdar, *Student Member, IEEE*, Patrick Benavidez, *Member, IEEE*, Paul Rad, *Member, IEEE*, Kim-Kwang Raymond Choo, *Senior Member, IEEE*, and Mo Jamshidi, *Fellow, IEEE*

***Abstract*— The aptitude to identify the emotional states of others and response to exposed emotions is an important aspect of human social intelligence. Robots are expected to be prevalent in society to assist humans in various tasks. Human–robot interaction (HRI) is of critical importance in the assistive robotics sector. Smart digital assistants and assistive robots fail quite often when a request is not well defined verbally. When the assistant fails to provide services as desired, the person may exhibit an emotional response such as anger or frustration through expressions in their face and voice. It is critical that robots understand not only the language, but also human psychology. A novel affection-based perception architecture for cooperative HRIs is studied in this paper, where the agent is expected to recognize human emotional states, thus encourages a natural bonding between the human and the robotic artifact. We propose a method to close the loop using measured emotions to grade HRIs. This metric will be used as a reward mechanism to adjust the assistant's behavior adaptively. Emotion levels from users are detected through vision and speech inputs processed by deep neural networks (NNs). Negative emotions exhibit a change in performance until the user is satisfied.**

***Index Terms*— Assistive robotics, human–machine interactions, humanoid robot, Internet of robotic things, smart home, supervisory control.**

I. Introduction

HUMAN–ROBOT interaction (HRI) has become more important in the last two decades as a result of people need to work closely with robots to achieve higher efficiency and productivity. Therefore, companies and government agencies are becoming interested more in robotic applications in such fields as the defense industry, energy, rehabilitation services, and home-user environments. In addition, human–robot cooperation tasks are growing in work environments. Through the advancement of artificial intelligence (AI) and the help of the growing literature, user-centered robotic applications are becoming a part of our daily lives. Autonomous robots are appearing in houses, on the streets and in offices. Robots have systems that can detect people, recognize voices and faces, and understand, learn, and perform repetitive work tasks. Autonomous vehicles, which arguably have seen the most drastic improvements recently, have systems that can recognize street signs, cars, and pedestrians, and determine the state of traffic lights. Other popular developments in robotic applications are varying, such as surveying landmarks from the sky, and helping everyday tasks in houses, the importance of these applications is founded upon the involvement of humans in the system.

The introduction of neural networks (NNs) has led to substantial increase in the performance of any intelligent system. Using Convolutional NN for images has provided us with well-trained systems that are able to detect and recognize faces, identify objects with high degree of accuracy and perform navigation tasks [1]–[6]. By the help of faster real-time tracking algorithms [7], the robots can infer the dynamics of the environment much easier, and perform accordingly and responsively, which is expected in a supervisory human–machine interaction use case [8], [9].

Evaluating the performance of any system that focuses on human–robot collaboration is the bottle-neck for HRI studies. Most of the time, well-explained metrics from the literature and statistical values of the system outcomes are calculated, and then the performance of the entire system is evaluated based on these numbers [10]–[12]; however, these metrics and methods have been losing their importance in recent years [13]. Nowadays, the interactions are more active than ever; robots are deployed with more advanced sensors, have vision capabilities with video cameras, listen to their environments, are connected to the local network, and have access to the Internet [14]–[16].

Therefore, the criteria for evaluating a system requires more interaction, listening and observing the users'

Manuscript received December 31, 2018; revised May 11, 2019; accepted June 10, 2019. This work was supported in part by the Air Force Research Laboratory and the OSD under Grant FA8750-15-2-0116. *(Berat A. Erol and Abhijit Majumdar contributed equally to this work.) (Corresponding author: Kim-Kwang Raymond Choo.)*

B. A. Erol, A. Majumdar, P. Benavidez, and M. Jamshidi are with the Autonomous Control Engineering Laboratories, Department of Electrical and Computer Engineering, The University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249 USA (e-mail: baerol@gmail.com; abhijit.g.majumdar@gmail.com; patrick.benavidez@utsa.edu; mo.jamshidi@utsa.edu).

P. Rad and K.-K. R. Choo are with the Department of Information Systems and Cyber Security, The University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249 USA (e-mail: peyman.najafirad@utsa.edu; raymond.choo@fulbrightmail.org).

Digital Object Identifier 10.1109/TCSS.2019.2922593

2329-924X © 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

feedback [17], and understanding their satisfaction levels in different ways, such as facial expressions. Humans are capable of simultaneously exhibiting a range of skills as a part of the communications, such as facial expressions that mimic our emotions, and gestures that represent short verbal statements. This multimodality increases the complexity of the communications [18] and the challenges in HRIs.

Humanoid robots are a well-examined class of robots potential to assist with routine repetitive tasks and daily activities in a household setting. Humanoid robots can observe, interact with, and imitate humans. They are performance policies are mostly built upon mimicking human functionality. As a result of recent advancement in deep learning algorithms and sensory devices, humanoid robots have the ability to perceive the world around them. The main roadblocks to incorporate humanoid robots into the work and home environments are the complexity of the unknown system dynamics under rare circumstances, safety, ethics, and the high costs of system components.

In this paper, experimental validation of emotion detection and recognition approach for improving the multi-modality in HRI is proposed. The role of the proposed system is to control the emotional responses created by ambiguities in the process of verbal and nonverbal exchanges between a robot assistant and an individual. The main scope of the study is improving the personalization in social robotics toward the users' emotional states.

The remainder of the paper proceeds as follows. Section II includes related works from the literature. Section III details the proposed HRI smart environment system, followed by Section IV, which provides a background on the literature of ambiguities of speech, facial emotional expressions regarding the multimodal approach. Section V describes the preliminary work that was performed for recursive emotion analysis experiments. Section VI details the extended experimental setup, and results based on the previous outcomes to reinforce the proposed system. Finally, Section VII presents the conclusions.

II. Related Works

Research into emotion detection has provided several useful tools and frameworks. Most of these frameworks used the parameters based on the facial action coding system (FACS) proposed by [19]. Most of the works in the literature used the feature points [20], [21] for distinguishing the facial expression recognition, where the variation of muscle distances was constructed based on the neutral emotional state. On the other hand, there are quite a number of approaches used the facial animation parameters (FAP), which represent a complete set of facial actions along with eye and lips motions and the orientation of the head [22]. Active appearance model, action units, support vector machines, and Gabor filters have been used for features extraction of emotion recognition processes [23]–[26].

Most of the approaches include a dynamic convergence property over time. For instance, for multimodal learning, the authors in [27] explained the importance of face, gesture, body language, speech or physiological signals on cognitive science for enhancing human-robot communication. Jaimes and Sebe [28] performed an extensive survey of multimodal emotion recognition approaches detailing human behavior for emotion recognition and human–computer interaction (HCI). In a separate work, Sebe *et al.* [29] used the feature points by analyzing displacement of varying facial muscles, and Bailenson *et al.* [30] also used feature points for calculating the facial contour deformations by taking the neutral state as a base. Bartlett *et al.* [31], on the other hand, combined the feature points selection and action units for facial expression recognition by using machine learning tools. Rabiei and Gasparetto [32] focused on determining the human emotion from voice and facial expressions, by attempting to extract a single emotional state from a set of possible emotions that would be a usable format for HRI. Faria *et al.* [33] attempted to determine emotional states from facial expressions for input into a small humanoid robot. Fig. 2 shows labeled sample inputs requested from a user for the most used emotional states suggested in the literature.

However, there are a number of limitations in the current literature. Most of the presented approaches used the same data sets and considered the neutral state as a baseline. Every emotional expression has complex muscle movement, which usually requires a specific classification. The existing approaches tend to assume that image sequence was perfectly partitioned, and users were starting from a neutral state and ending on highest state for the expressed emotion. Unfortunately, this is not indicative of a real-world scenario. It is not always true that the development of expression strictly follows the state of neutral, the expression, and then, neutral again. One's particular daily mood, or a dominant mood for a particular time of the day can play an important role while constructing a baseline. Therefore, there is a need for discretizing the transitions between the emotional states while allowing a rapid classification.

III. Proposed Emotion-Based Assistive System for Active Human–Robot Interactions

The proposed environment for assistive robotic systems includes audio and graphical interfaces, a cloud service connection and a mobile robotic platform, as shown in Fig. 1. In this section, we discussed hardware preparation, emotional state detection, and the control loop.

A. Robotic System Prototype

For the assistive system, we built a hybrid design, a humanoid robot, and a Kobuki Turtlebot 2 from Yujin-Robotas, as shown in Fig. 1. For the humanoid component, a 3-D printed open-source robot, Poppy Torso [35], was modified and used. The robotic platform is compatible with the Robot Operating System (ROS). We used a camera on the humanoid's head for object detection tasks. The open-sourced *kobuki ROS package* controls the mobile platform. The modified head unit holds ODROID-VU5 touchscreen liquid crystal display (LCD), a monocular camera, a set of stereo speakers, and a microphone. A software tool was developed by using pyqt3 for the graphics. A ROS software package *ace_poppy_hri_display* from [36] was used for this work. A question and answer game was followed by the robot and user of the system. The camera

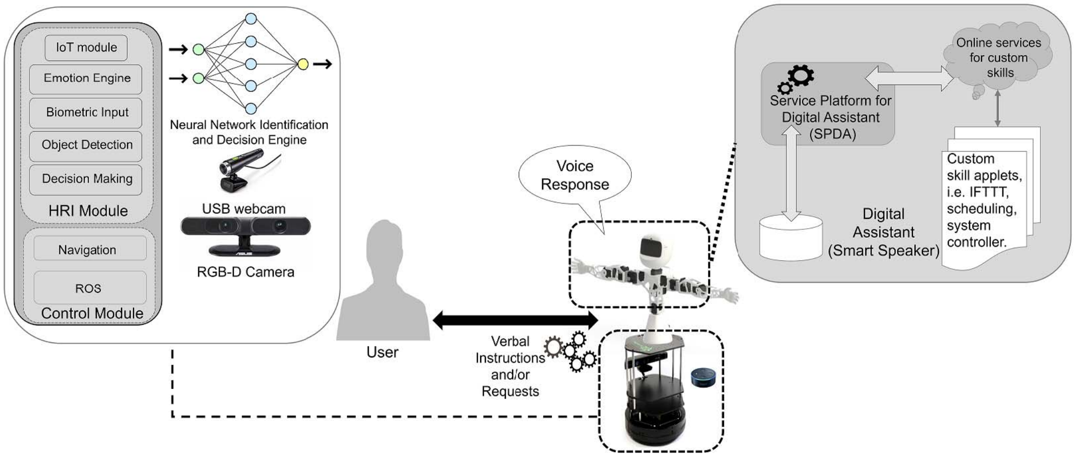

IoT module
Emotion Engine
Biometric Input
Object Detection
Decision Making
HRI Module
Navigation
ROS
Control Module
Neural Network Identification and Decision Engine
USB webcam
RGB-D Camera
User
Verbal Instructions and/or Requests
Voice Response
Service Platform for Digital Assistant (SPDA)
Online services for custom skills
Custom skill applets, i.e. IFTTT, scheduling, system controller.
Digital Assistant (Smart Speaker)

Fig. 1. User-centered approach for the system testbed of the robot assistant for emotion detection. It is reinforced with a smart digital assistant for voice activation, and an IoT module focused on high-level HRIs.

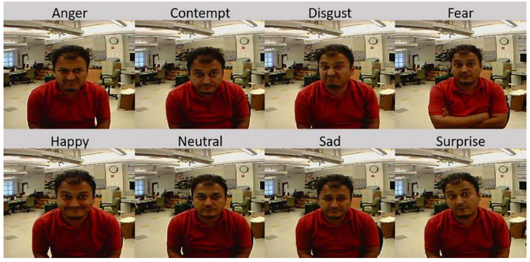


Fig. 2. Examples of inputs as suggested by the literature and with the highest score to the related emotional expressions from [34].

is used as a source of data for detecting emotions. Images from the camera are acquired through ROS using the software package *uvc_camera*.

B. Auditory Interface

Auditory commands are gathered from a digital voice assistant device. User's vocal responses to the humanoid robot are then processed through a combination of programming applications, *espeak* and *aplay* for the synthesis of words into a WAV file. One of the reasons to use a WAV file is the tendency of *espeak* to connect to the jack-server slowly; then, it fails intermittently, whereas *aplay* plays back the audio almost instantaneously and consistently. *espeak* is also known to be limited to pronounce certain words and names. Another reason is to maintain the responses to the user for further training purposes.

C. Control Loop

The control loop is initiated by a new task request from the user. A digital assistant with cloud service connection fetches the requested action and schedules an action plan to the robot. Until the robot fully accomplishes the task of searching for the appropriate solution, regular status updates are provided to the user. Once the robot completes the task, results are provided to the user to determine if the actions were correct. In the case of ambiguities of the user request, the presented information can be an array of possibilities that the system came up with. Emotional responses graded by the system indicate the appropriateness of the response to the user's desired actions of the system. If the response is correct, the robot finishes any other components of its task.

There are multiple layers of complexity in controlling a system with emotion feedback. The features described in the sections above contribute to the making of such a system, and the control of each aspect of the system is essential for the proper execution of different tasks in this HRI system. This is performed by using a text-to-voice translation of query scripts. The emotion of the user is analyzed by recording their expression using the camera modification on the robot and then passing it through the modified emotion analysis network which will be described in the following section.

IV. Multimodal Emotion Detection for System Personalization in HRI

The aptitude to identify the emotional states of others and respond to exposed emotions is an important aspect of human social intelligence. HRI is a critical component in the home and assistive robotics sector. Various facial expression recognition systems and frameworks are presented in the literature for an effective HRI. Most of them try to recognize emotional responses through a social dialog by gathering a vocal and visual data from the users. This multimodality increases the complexity of the communication and the challenges in HRIs.

To use emotion to close the loop in HRI cases, the solution would include a question and answer type along with history. The answering portion of the process can be verbal or visual. The system's performance can be evaluated based on the vocal response and facial expressions of the user, which are signifying the overall user satisfaction level. Smart digital assistants and assistive robots fail quite often when a request is not well defined verbally. In such cases, the user may exhibit an emotional response such as anger or frustration through expressions in their face or voice. A history of answers along with emotional responses can be recorded to best suit preferences of the user in the future [37]. For HRI scope, in addition to the vocal response, the system performance can

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE I

DISTRIBUTION OF CONFIDENCE SCORES FOR EACH EMOTIONAL STATE OBTAINED FROM [34]. TWELVE TRIALS WERE PROCESSED ON THE API, TWELVE PICTURES OF THE USER WERE UPLOADED FOR EACH STATE, AND THE AVERAGE SCORE WAS CALCULATED. THE TOP TWO SCORES ARE IDENTIFIED FOR EACH ROW IN THE TABLE. AS SEEN FROM THE TABLE, *Happiness*, AND *Neutral* STATES DOMINATE THE CONFIDENCE LEVEL SCORES, WHEREAS THE RESTS ARE WEAKLY PREDICTED

	Scores from Emotion API							
Expressed Emotion	Anger	Contempt	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	**0.3996**	0.0389	0.1224	0.0011	0.0008	**0.4235**	0.0059	0.0078
Contempt	0.0000	0.0541	0.0001	0.0000	**0.0553**	**0.8685**	0.0090	0.0128
Disgust	0.0970	0.0568	**0.1348**	0.0053	0.1718	**0.3918**	0.1052	0.0371
Fear	0.0080	0.0113	0.0812	0.1372	0.0369	**0.1896**	0.0353	**0.4985**
Happiness	0.0001	0.0003	0.0005	0.0006	**0.9355**	**0.0124**	0.0003	**0.0502**
Neutral	0.0001	**0.0137**	0.0001	0.0000	0.0019	**0.9661**	0.0103	0.0077
Sadness	0.0002	0.0701	0.0091	0.0006	0.0043	**0.1556**	**0.7590**	0.0011
Surprise	0.0001	0.0017	0.0002	0.0062	**0.1220**	0.0143	0.0000	**0.8554**

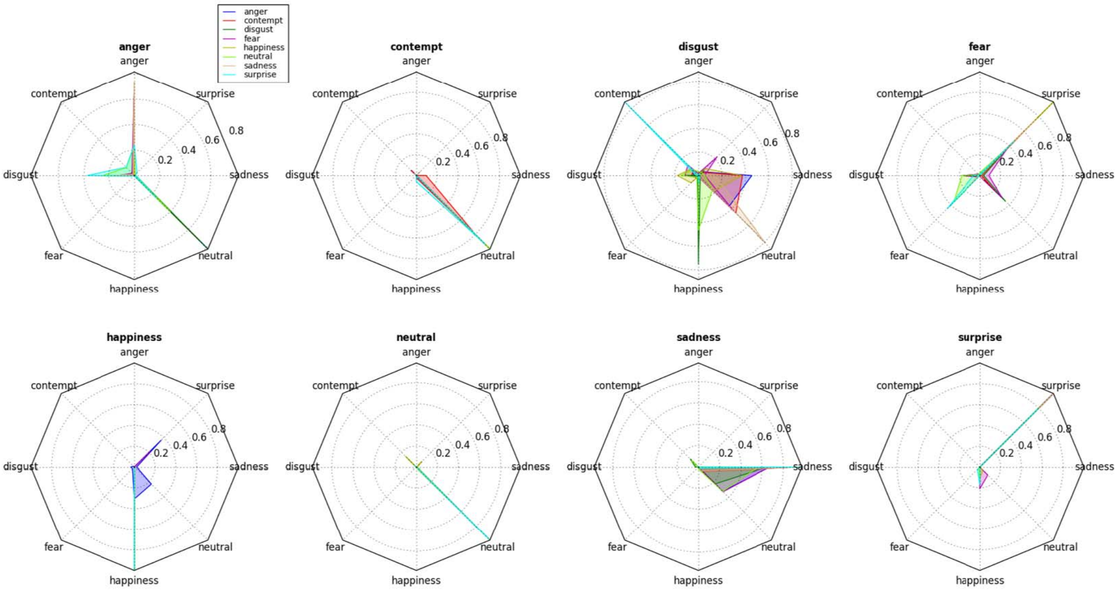


Fig. 3. Comparison of the expressed emotions to calculated emotions for the User #1 detected by Microsoft Emotion API [34]. The emotion predictions, confidence scores are mostly resulted as expected, except partial differences on *contempt*, *disgust*, and *fear* states.

be mapped into the user's facial expressions that represent their current emotional states. An example output for a single face from [34] for happiness is provided below.

```
Face Location Rectangle:
    "top": 114,
    "left": 212,
    "height": 65
    "width": 65
Emotion Scores :
    "happiness": 0.9999998,
    "sadness": 1.23025981E-08,
    "anger": 1.0570484E-08,
    "fear": 6.00660363E-12,
    "surprise": 9.91396E-10
    "disgust": 1.60232943E-07,
    "contempt": 1.52679547E-09,
    "neutral": 9.449728E-09,
```

Relatively close confidence scores have been calculated for different emotion states for one expressed emotion. We found that Microsoft Emotion application programming interface (API) provides high confidence scores especially for *happiness* and *neutral* states regardless of the emotion expressed by the user. Unfortunately, it lacks in scoring any other expressed emotional state with a lower confidence level by providing unreliable scores. Table I raises the concerns on unreliable confidence scores for each emotional states. Furthermore, Figs. 3 and 4 illustrate overall scores of expressed emotions, and sadly support those concerns. They show example output of emotion states for two users with different skin color from different nationalities mapped into a radar plot to show the relative scores gained from [34].

As seen in Fig. 4, it is interesting to note that an expressed emotion of *sadness* did not map directly to a large increase in the *sadness* score for another user. Instead, the scores for the *neutral* and *happiness* were sharing the highest confidence level scores, which repeated for both *angry* and *happiness* scores as well; comparing from Fig. 3, these emotion states have relatively unbalanced confidence levels. After processing these two emotional state samples from the aforementioned

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

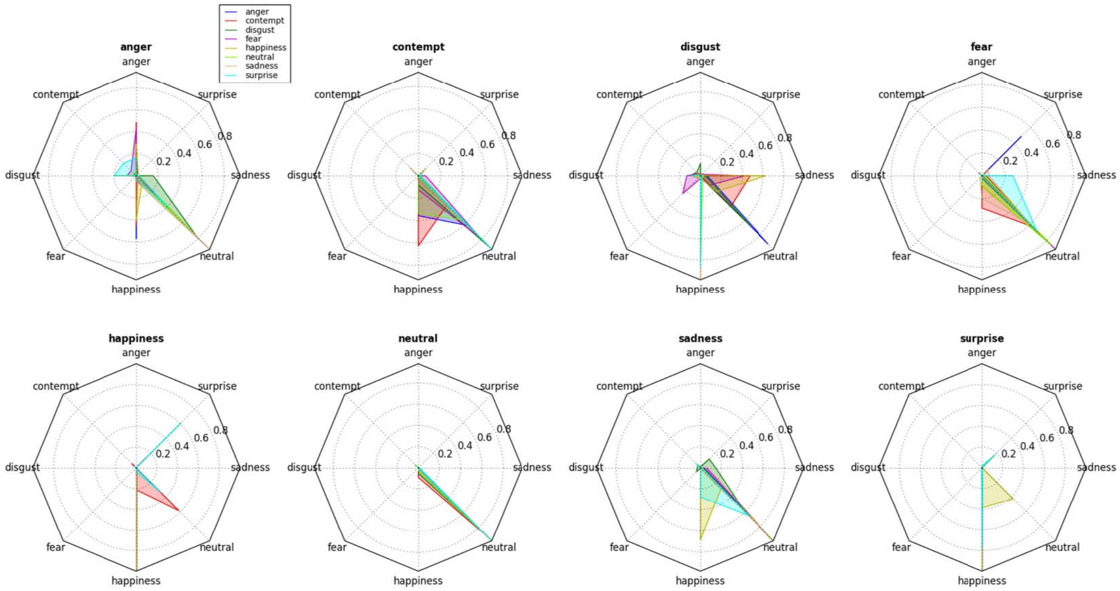


Fig. 4. Comparison of the same expressed emotions to calculated emotions for the User #4 detected by [34].

experiments, we concluded that the confidence scores for these emotional states are not reliable, and provide biased scores based on the color of the user's skin. On the other hand, these emotional states can be the best evaluation criteria for preventing any ambiguity scenarios in conversation-based interactions between human users and robots in HRI studies. We found that a recent study states a research outcome similar to our concerns, based on facial recognition and gender detection tools and their biased databases [38]. Based on the output from earlier experiments, and guidance from the literature, the weighted calculation should be more selective. Under these concerns, we have decided to extend the number of users from various nationalities and races to perform similar emotional facial states to test this conclusion and bring about a more robust outcome for HRI scenarios.

The set of emotions reported in the literature, such as anger, disgust, happiness, neutral, sadness, and surprise, can be used to gauge the satisfaction level of a person relating to human–robot collaborations. Due to biased confidence level scoring for emotional states, the accuracy of the scores in [34] is debatable. Therefore, an error theoretically can be calculated as a weighted summation of the score output from an emotion detecting network

$$e = \sum_{i=0}^{N-1} w_i s_i \tag{1}$$

where N is the number of emotions, i is an iterator though an enumeration of emotion score names, w_i is the weight for score i, and s_i is the emotion score value for score i. With the summation expanded out and using emotion score names instead of the enumerated values, e would be the following:

$$\begin{aligned} e = {} & w_A s_A + w_C s_C + w_D s_D + w_F s_F \\ & + w_H s_H + w_N s_N + w_S s_S + w_{\mathrm{Su}} s_{\mathrm{Su}}. \end{aligned} \tag{2}$$

Weight values will be determined based on initial training to the user's expressions when prompted to feel the various emotions; such as *A*nger, *C*ontempt, *D*isgust, *F*ear, *H*appiness, *N*eutral, *S*adness, and *Su*rprise, subsequently through general conversation with the robot.

V. Recursive Emotion Analysis

Convolutional NNs (CNNs) are leading a state-of-the-art approach for facial expression and emotion recognition research. Current literature on emotion recognition depends on the same image data sets. Increasing the diversity and the volume of the data set will not only increase the accuracy of the new methods but also will improve the overall system performance. For bounding an effective HRI, the literature suggests several studies that evaluate the advantage of using facial emotional expressions. Most of these studies are based on imitating different human behaviors and actions. It is interesting to know that the literature suggests that the previous works have weakly distinguished the anger and disgust emotional states, and frequently confused of the sadness and happiness with anger. When it comes to emotion recognition, neutral, surprise, and happiness were the easiest to recognize.

Emotion detection algorithms developed using deep NNs provide a probabilistic model to show the most likely emotion that the user is expressing [39]–[42]. Such algorithms use each frame from the input to determine the output vector representing probabilities of different *microemotions*, meaning that they represent the current emotion of the user as interpreted by the computer. There are challenges associated with using this data to analyze a user's emotion and make decisions based on them. The output probabilities are fluctuating in a continuous sequence of frames that the camera on a robot may capture, and resulting in a noisy prediction. One may argue for using any particular frame from this sequence to determine the emotion; however, expressions of emotions are subjective to the user. Also, the microexpression obtained from a single frame from the user is not a true representation of the emotional state of the user.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

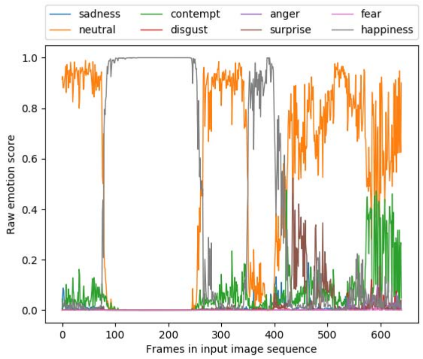


Fig. 5. Fluctuating microexpressions prediction plot on input image sequence over 20 s.

A plot of the results of emotion analysis obtained from a sequence of frames of a user is shown in Fig. 5. Emotions expressed by the user do not change at such high frequency. For instance, while having a conversation with a person, one's opinion might be expressed as a facial expression when a couple of arguments or statements have been made, which cannot be judged by analyzing emotion in one single picture of the person. Both of these issues may be addressed by considering the user input from the camera as a temporal sequence of frames, and thus considering the *macroexpressions* of the user. Our hypothesis is that using such a sequence of input images, the change in the user's emotion is more informative in making a prediction on whether the user is satisfied with the robot or not. This method also addresses the noise in the prediction data, since considering a temporal sequence of data may filter out the noise and provide a better approximation of the result. It also takes into consideration the fact that each individual is analyzed differently by the emotion API. For instance, if someone is already happy, their expression to agree with an argument might be the user being happier. In such cases, the neutral level of the person may be different even though his current expression might be set to a value other than neutral, and the change in the expression is important to determine the positive or negative reaction of the user.

A. Long Short-Term Memories

Long Short-Term Memories (LSTM) by Hochreiter *et al.* [43], was one of the most successful solutions to the problem of *vanishing gradient descent*, inherent to traditional Recurrent NNs (RNNs), which confuses the recurrent networks to learn long term dependencies of a sequence. In the case of emotion analysis, we need to consider the relative change in their emotions as explained earlier. Thus, an LSTM network is desirable, as it enables the network to learn the user's emotion features. An LSTM layer is made up of LSTM cells, each of which contains gated memory blocks, along with the cell state. These gates are modified with training to learn to either retain or forget part of the previous cell state, considering the current input. These features introduced later are key to the success of LSTM for long-term dependencies [44], [45]. The following equations provide the LSTM for the forward pass and the backward pass for a single LSTM cell.

1) Forward Pass: Input gates:

$$\begin{aligned} a_\iota^t &= \sum_{i=1}^{I} w_{i\iota} x_i^t + \sum_{h=1}^{H} w_{h\iota} b_h^{t-1} + \sum_{c=1}^{C} w_{c\iota} s_c^{t-1} \\ b_\iota^t &= f(a_\iota^t) \end{aligned} \tag{3}$$

Forget gates:

$$\begin{aligned} a_\phi^t &= \sum_{i=1}^{I} w_{i\phi} x_i^t + \sum_{h=1}^{H} w_{h\phi} b_h^{t-1} + \sum_{c=1}^{C} w_{c\phi} s_c^{t-1} \\ b_\phi^t &= f(a_\phi^t) \end{aligned} \tag{4}$$

Cells:

$$\begin{aligned} a_c^t &= \sum_{i=1}^{I} w_{\mathrm{ic}} x_i^t + \sum_{h=1}^{H} w_{\mathrm{hc}} b_h^{t-1} \\ s_c^t &= b_\phi^t s_c^{t-1} + b_\iota^t g(a_c^t) \end{aligned} \tag{5}$$

Output gates:

$$\begin{aligned} a_\omega^t &= \sum_{i=1}^{I} w_{i\omega} x_i^t + \sum_{h=1}^{H} w_{h\omega} b_h^{t-1} + \sum_{c=1}^{C} w_{c\omega} s_c^{t-1} \\ b_\omega^t &= f(a_\omega^t) \end{aligned} \tag{6}$$

Cell Outputs:

$$b_c^t = b_\omega^t h(s_c^t) \tag{7}$$

2) Backward Pass:

$$\epsilon_c^t = \frac{\delta O}{\delta b_c^t}, \epsilon_s^t = \frac{\delta O}{\delta s_c^t} \tag{8}$$

Cell Outputs:

$$\epsilon_c^t = \sum_{k=1}^{K} w_{\mathrm{ck}} \delta_k^t + \sum_{h=1}^{H} w_{\mathrm{ch}} \delta_h^{t+1} \tag{9}$$

Output Gates:

$$\delta_w^t = f'(a_w^t) \sum_{c=1}^{C} h(s_c^t) \epsilon_c^t \tag{10}$$

States:

$$\epsilon_s^t = b_w^t h'(s_c^t) \epsilon_c^t + b_\phi^{t+1} \epsilon_s^{t+1} + w_{c\iota} \delta_\iota^{t+1} + w_{c\phi} \delta_\phi^{t+1} + w_{\mathrm{cw}} \delta_w^t \tag{11}$$

Cell:

$$\delta_c^t = b_\iota^t g'(a_c^t) \epsilon_s^t \tag{12}$$

Forget Gates

$$\delta_\phi^t = f'(a_\phi^t) \sum_{c=1}^{C} s_c^{t-1} \epsilon_s^t \tag{13}$$

Input Gates

$$\delta_\iota^t = f'(a_\iota^t) \sum_{c=1}^{C} g(a_c^t) \epsilon_s^t \tag{14}$$

where w_{ij} is the weight of the connection from unit i to unit j, a_i^t is the network input to unit j at time t, b_i^t is the value of the

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

same unit after the activation function has been applied. ι is the input gate, and ϕ is the forget gate, while ω is the output gate, and C is the set of memory cells of the block. s_c^t is the state of cell c at time t, whereas f is the activation function of the gates; here, $f(x) = (1/1 + e^{-x})$, and g, the cell input activation function, is $g(x) = (4/1 + e^{-x})$. h is the cell output of activation function, $h(x) = (2/1 + e^{-x}) - 1$. O, objective function, is used for training, and I, number of inputs, K, number of outputs, H, number of cells, are in hidden layer.

B. Emotion Sequence Analysis Using LSTM

The use of LSTM layers on a sequence of frames to analyze emotion helps in determining the transitions in emotions over a temporal constraint. The LSTM layer is extended after the CNN in the form of transfer learning, where the trained CNN network is fixed while training the LSTM layer on a different data set. As a result, since the whole architecture is not being trained, a comparatively smaller data set is needed to train the LSTM layer.

The LSTM used for our experiment consists of a single layer of 64 LSTM cells, the output of which is densely connected to the output. The input to the LSTM network is the output obtained from a generic CNN-based emotion analysis architecture, which provides a vector of 8 different emotions, representing the probabilities of each emotion being expressed by the user captured by the input frame. Since we are concerned with a sequence of images, through a live video feed of the user, the output vector of the emotion analysis of 48 consecutive image frames are fed as the input to the LSTM network. Hence, the LSTM layer has 48 time-steps to iterate over, with 8 input features each, generating an output of 64 features, which are passed through a fully connected layer to generate an output vector $\vec{E}_{\text{prediction}}$ with a dimension of 3 corresponding to the emotion transitions *Toward happy*, *Toward sadness*, and *Toward neutral*. These three transitions are then used to determine the user response to an action taken by the robot. The general architecture of the network is shown in Fig. 6. A summary of the hyper parameters that define its architecture and used to train the LSTM network is shown in Table II.

To train the model, we created a data set by recording several image sequences while expressing different emotions and then passing these emotions through the emotion analysis framework, to obtain a data set of a constant sequence of emotion vectors. Each of these emotion vectors contain the level of sadness, neutrality, contempt, disgust, anger, surprise, fear and happiness expressed by the user. The user is asked to change their expression of emotions while the image sequences are recorded. Thus, the data set contains a time-dependent sequence of inputs, which can be fed to the LSTM network. The network was trained on a system with an Intel Core i5-6600 at 3.30-GHz quadcore processor with 8-GB memory on a Nvidia GTX1080 graphics processor to facilitate expedited learning from different users.

VI. Experimental Results

In this section, we are introducing the experimental results and discussions from a real-time application that enhances the robotic system to understand the emotional state of a user. This metric can be used to adjust the system's behavior adaptively. A novel affection-based perception architecture for HRIs is presented. The robot is expected to recognize human emotional states, thus encouraging a natural bonding between the human and the robotic artifact. We present a method to close the loop using measured emotions to grade the effectiveness of HRIs. Negative emotions will exhibit a change in performance until the user is satisfied.

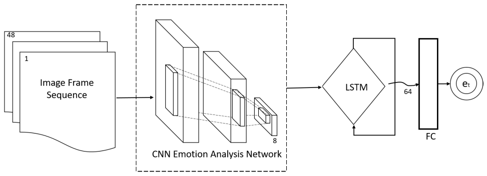


Fig. 6. Modified NN architecture used for temporal analysis of macroemotions.

TABLE II
Hyperparameters Used for LSTM Based Deep Network

Model			
Layers	Input	Output	Activation
LSTM layer	48(timesteps)x8(features)	64(cells)	None
Dense layer	64 (neurons)	3 (output)	Softmax
Training			
Loss	Categorical crossentropy		
Optimizer	RMSProp		
α	0.001		
ρ	0.9		
ϵ	None		
Decay	0.0		
Epoch	5000		
Validation split	10%		
Batch size	400		

A. Emotion Sequence Analysis

Different image sequences were recorded from users, who were asked to express opinions transitioning from neutral to either happy or sad, in order to train and test the network. For the purpose of training the network, we used image frame series of transitions from neutral to happy as approval, transitions from neutral to sad as disapproval, and all other series of transitions to different emotions as neutral. The data set we used was required to be pre-processed to ensure that the initial neutral state of the user was not labeled as transitions to happiness and sadness. Fig. 7 shows some of the results of the processed data and the raw emotions acquired from the transitions in the testing data set. For instance, in Fig. 7 (a), the raw micro-emotions in the image frame show that the user expressed sadness for a short time, and confidence score for sadness is higher then neutral state during that time (blue dashed line). Then, overall confidence score is turning back to the neutral again (orange dashed line). This would be due to the user being neutral for the system response; however, user's expression in the experiment was sadness. This bring us to the level of emotion user exhibits which is subjective; therefore, the system is trained to the user's emotions, it learns in order to predict the correct emotion transition (solid black line).

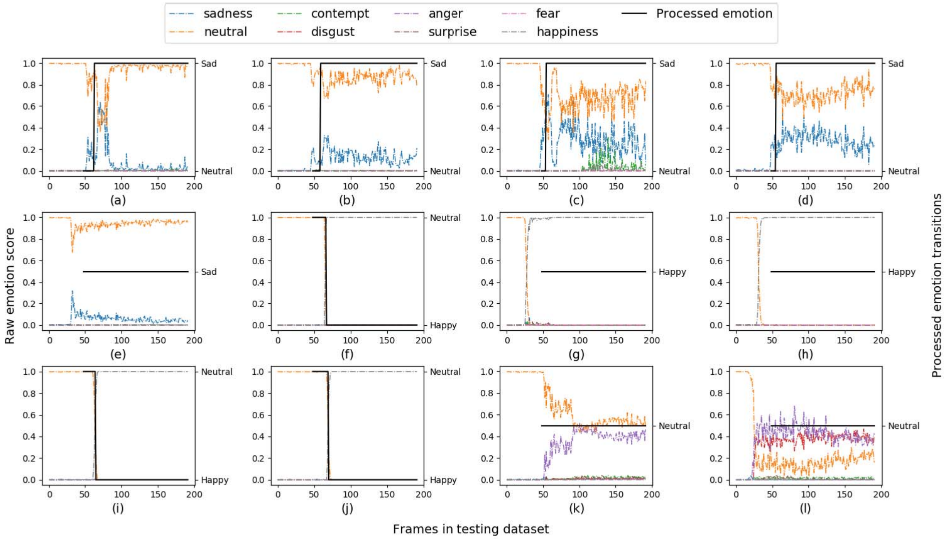


Fig. 7. Macroemotion transition detection on testing data set using trained network. (a)–(e) Prediction on transitions to sadness. (f)–(j) Prediction on transitions to happiness (gray dashed line), the easiest emotional state to predict as suggested in the literature, our approach quickly captures the happiness as well. (k)–(l) Prediction on other emotion transitions, the network trained to be predict them as neutral.

It can be observed that the use of a temporal constraint at the output clearly distinguishes between the different emotions expressed by the user, which can now be used to determine the positive or negative response of the user. This is a good candidate to overcome the complexity of determining the emotional state changes, as illustrated in Fig. 5. It should also be noted that the network was trained to perform temporal analysis to transition to seven other emotions from neutral. For determining the user's approval of the system outcome (i.e., evaluating the user satisfaction level on the response), only the transitions to happiness and sadness were considered.

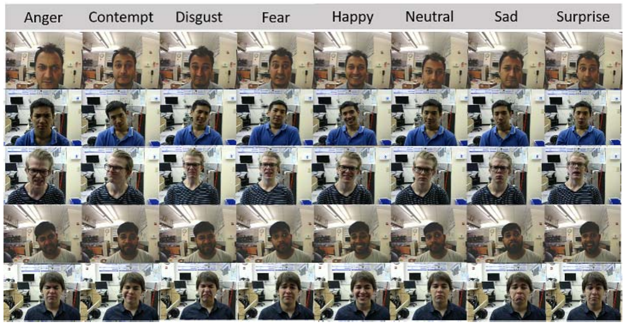


Fig. 8. New group of users from various races and nationalities, and they are distinguished by their skin colors.

B. Discussion

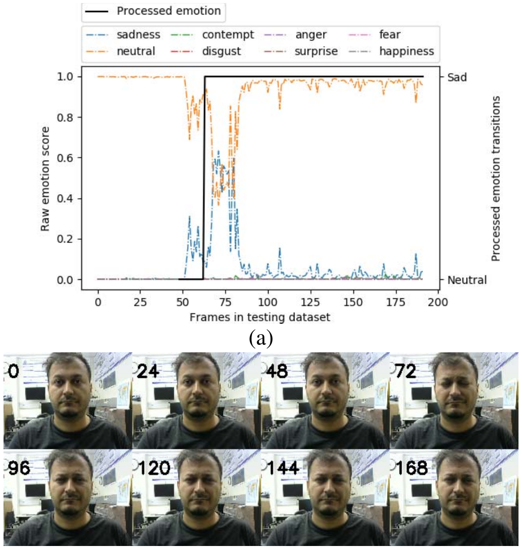
As stated in Section III, because of weak confidence level scoring and biased databases for face detection and emotion recognition tools for different skin colored users, these outcomes increase the concerns on implementations of available approaches in HRIs. The preliminary results show that our approach could be a good fit to remove any of these concerns. Although our system included a limited number of users in the beginning due to assistive robotic work environments, we have decided to extend the number of users and their profiles to overcome these issues as shown in Fig. 8.

During our studies, we found questionable, and some may even say biased, results, which suggest the Microsoft emotion API can be selective in providing lower confidence scores for particular emotion states, such as *happiness* and *sadness*, based on the user's skin color. These emotion states are crucial for any vision-based HRI study, since they will be mimicking the user's experience level from the system's performance, such as satisfaction or dissatisfaction levels, and building a metric for the system's reliability and efficiency. For this purpose, we created a data set that includes 22 464 pictures of 5 individuals, in which each individual was asked to express relevant emotional state, and the pictures were tagged based on the expressed emotions. The database is built upon eight different emotional state, and several trials conducted for a sequence of pictures of users that expressing the related emotion. Therefore, to train the system, we parsed a picture that includes 64 frames where the user was showing a unique way of their expression of requested emotion. This was processed for each emotional state; then, we repeated the same cycle for recursive emotional analysis, as stated in Section IV. Furthermore, we train the network for specific emotional expressions as well as the emotional transitions for understanding the satisfaction and dissatisfaction boundaries; then, the data set was updated.

Furthermore, we have found that the API even had complexity in recognizing faces. While we were processing the user's pictures, the number of faces detected by the API was dramatically decreased. Table III represents the accuracy of the face detection process of a trial in the extended experiments. As shown in the Tables IV, V and VI, the results for each emotional state for each user with the same expression were weakly predicted, and the scores were radically different, one may even say this was due to irrelevant confidence levels based on user's skin color. Once again, the scoring for the expressed emotions shows huge differences based on the users' facial

TABLE III
NUMBER OF DETECTED FACES AND THE ACCURACY RATES

	No. of faces detected	Total no. of pictures	Accuracy
User #2	772	1536	0.5026
User #3	762	1536	0.4961
User #5	1536	1536	1.00
Total	3070	4608	0.6662



(a)

(b)

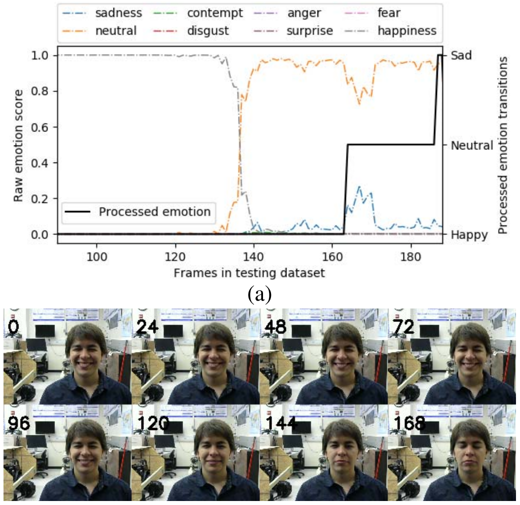
Fig. 9. Temporal constraint macro emotion transition. (a) Emotion transition. (b) Image sequence used to determine emotion transition.

skin colors, the scores for darker skin are highly *happiness* and *neutral* oriented.

The experimental results obtained after training the network proves our hypothesis, as can be clearly seen in Fig. 9. The raw microemotions in the image sequence show that the user only expresses sadness, interpreted as disapproval, for a fraction of the time, while being neutral for the rest of the series. However, it can be clearly seen from Fig. 9(b), that the user transitions from neutral to sad and stays sad throughout the series. This characteristic of the level of emotion a person exhibits is subjective; and thus, the network was trained to the user's emotions, which it then learns in order to predict the correct emotion transition as seen in the Fig. 9(a). It should also be noted that using the highest predicted microemotion directly would only show a neutral emotion, which is both incorrect since the subject was asked to express sadness, as well as unusable in determining the correctness of the task performed by the robot.

In the control loop, we used this expresses to judging the user's reaction when asked a question or when an object is queried. The camera on the robot starts to capture the user expressions while the question is being asked. Furthermore, it is important to state that the LSTM layer in the network can only make a prediction once the initial 48 frames are acquired. Thus, some of the test series observed in Fig. 7 predict the transition state immediately. As a result, the transition to approval or disapproval by emotion is observed.

Our results become even more important in this stage, in such situations where the user may not *ideally* express certain emotions. For example, one such user that we performed our tests on, is usually a *happy* person. This means that even when he expresses sadness, the resulting expression calculated by microexpression analysis may only indicate a neutral emotion, as can be observed in the test results shown in Fig. 10. In the figure, as the subject transitions from happy to sad, the micro expression scores indicate a likelihood of a neutral expression by the user. Since our algorithm is subject specific, it identifies the nuances in the sadness emotion scores, along with the neutral score to indicate a transition to sadness by the user, as observed in Fig. 10(a).



(a)

(b)

Fig. 10. Temporal constraint macro emotion transition from happy to sad. (a) Emotion transition. (b) Image sequence used to determine emotion transition.

For instance, there were trials that the Microsoft API was not able to recognize some of the faces in the same sequence of the picture frames and there were times that this number even turned out zero. Therefore, this has resulted in the exclusion of those frames for comparing our results fairly. Since most of the recent works in the literature were using off-the-shelf tools, and the data sets they used became commercial product in a short time, we have compared our results with the [34], which is often used by those research works and referred as a baseline in the literature. As shown in Fig. 11, the number of correctly predicted emotional states increases toward to the users with lighter skin color, showing the same behavior for the face recognition rate partially. The reason of receiving lesser emotion prediction for the lightest skin colored user is resulting on the weak face recognition rate for the same user. Moreover, it is easy to conclude from Fig. 12 that our system performs better for detecting the user's emotional expressions and shows higher accuracy rates for almost all emotional states, comparing 50.68% to 67.36%. For example, the average user accuracy for each emotional state of our system is 11% higher than that of [34]. The following charts in Fig. 11 represent the accuracy of face detection rate for the users, and the number of emotions correctly recognized by [34]. Furthermore, the normalized confusion table for each emotional expression by the users are shown in Fig. 12.

TABLE IV

DISTRIBUTION OF CONFIDENCE SCORES FOR EACH EMOTIONAL STATES OBTAINED FROM MICROSOFT EMOTION API IN EXTENDED STUDIES FOR USER #2. THE TOP TWO SCORES ARE IDENTIFIED FOR EACH ROW

Expressed Emotion	Anger	Contempt	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	**0.2503**	0.0129	0.0185	0.0004	0.0000	**0.4932**	0.2242	0.0005
Contempt	0.0045	**0.1372**	0.0040	0.0008	0.0153	**0.8190**	0.0126	0.0065
Disgust	0.0104	0.043	0.0203	0.0024	**0.7744**	**0.1213**	0.0642	0.0027
Fear	0.0001	0.0020	0.0000	0.0000	**0.0109**	**0.9859**	0.0009	0.0001
Happiness	0.0000	0.0000	0.0000	0.0000	**1.000**	0.0000	0.0000	0.0000
Neutral	0.0003	0.0053	0.0004	0.0000	**0.4098**	**0.5646**	0.0195	0.0000
Sadness	0.0010	0.0032	0.0004	0.0000	0.0000	**0.6584**	**0.3369**	0.0000
Surprise	0.0076	0.0008	0.0003	0.0051	0.0001	**0.8970**	0.0056	**0.0835**

TABLE V

DISTRIBUTION OF CONFIDENCE LEVEL SCORES FOR EMOTIONAL STATES OBTAINED FROM MICROSOFT EMOTION API FOR USER #3

Expressed Emotion	Anger	Contempt	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	0.0938	0.0175	0.1144	0.0056	**0.4457**	**0.2650**	0.0387	0.0193
Contempt	0.0116	0.0979	0.0119	0.0005	**0.2239**	**0.6400**	0.0105	0.0037
Disgust	0.0460	0.0059	**0.2889**	0.0076	0.0032	0.1614	**0.4842**	0.0027
Fear	0.0088	0.0015	0.0362	0.0031	**0.6633**	**0.1655**	0.0154	0.1063
Happiness	0.0003	0.0001	0.0010	0.0000	**0.9934**	**0.0050**	0.0001	0.0001
Neutral	0.0019	**0.0076**	0.0004	0.0000	0.0003	**0.9870**	0.0017	0.0011
Sadness	**0.1095**	0.0523	0.0423	0.0042	0.0009	**0.6822**	0.0994	0.0086
Surprise	0.0067	0.0015	0.0048	0.0625	0.0010	**0.5912**	0.0108	**0.3214**

TABLE VI

DISTRIBUTION OF CONFIDENCE LEVEL SCORES FOR EMOTIONAL STATES OBTAINED FROM MICROSOFT EMOTION API FOR USER #5

Expressed Emotion	Anger	Contempt	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	**0.2208**	0.0185	0.1911	0.0001	0.0027	**0.5322**	0.0344	0.0002
Contempt	0.0008	**0.4299**	0.0024	0.0000	**0.4256**	0.1367	0.0046	0.0000
Disgust	**0.2784**	0.0009	**0.6776**	0.0009	0.0021	0.0007	0.0393	0.0000
Fear	0.0000	0.0039	0.0002	0.0000	0.1349	**0.4014**	**0.4208**	0.0001
Happiness	0.0000	0.0000	0.0000	0.0000	**0.9999**	0.0000	0.0000	0.0000
Neutral	0.0004	0.0025	0.0007	0.0000	0.0003	**0.8729**	**0.1230**	0.0002
Sadness	0.0004	0.0315	0.0263	0.0037	0.0000	**0.0473**	**0.8901**	0.0008
Surprise	0.0005	0.00618	0.0026	0.0014	0.0309	**0.7924**	**0.0651**	0.0452

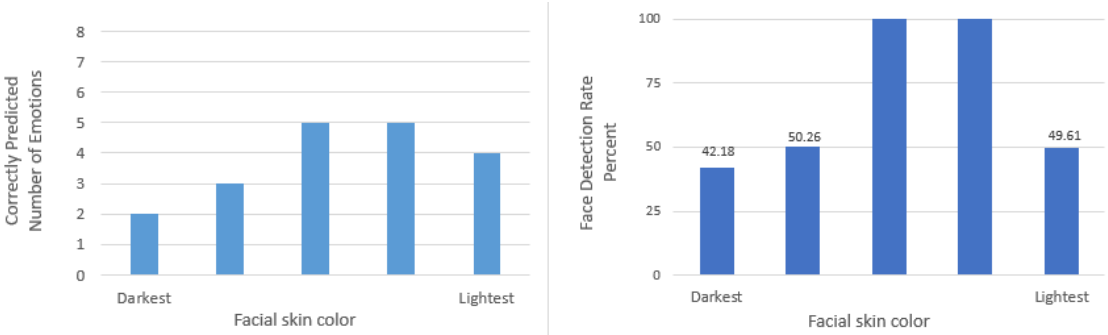
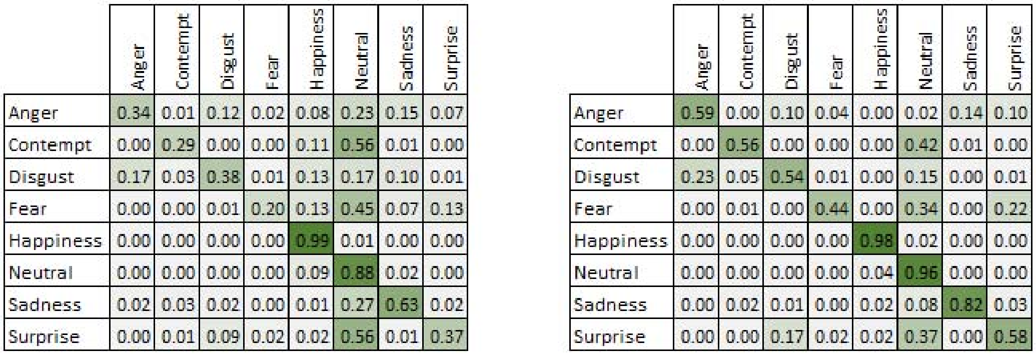


Fig. 11. Results illustrates the accuracy for emotion detection and face recognition rates for the users from [34]

C. Limitations

Although this work is targeted toward emotion analysis through a robot with a human in the loop scenarios, when using such inference to close a control loop system, one needs to consider the variability of a human's emotional state. This is highly subjective to several factors related to the nature of the person being observed, hence a much broader scope of data is required to train the network to account for a more diverse set of human emotional states.

As the system is intended to be used as a real-time application, a future research point is also to consider the use of a preprocessing network to take in images to produce emotional analysis on static images, which later are fed to the LSTM network defined in this research. This is limited to the availability of the extensive amount of labeled data and hence a current restriction for this research. The number of images used as a sequence is also determined arbitrarily and has a scope of improvement using empirical data. Images of a person's face may not always available in all situations;



	Anger	Contempt	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	0.34	0.01	0.12	0.02	0.08	0.23	0.15	0.07
Contempt	0.00	0.29	0.00	0.00	0.11	0.56	0.01	0.00
Disgust	0.17	0.03	0.38	0.01	0.13	0.17	0.10	0.01
Fear	0.00	0.00	0.01	0.20	0.13	0.45	0.07	0.13
Happiness	0.00	0.00	0.00	0.00	0.99	0.01	0.00	0.00
Neutral	0.00	0.00	0.00	0.00	0.09	0.88	0.02	0.00
Sadness	0.02	0.03	0.02	0.00	0.01	0.27	0.63	0.02
Surprise	0.00	0.01	0.09	0.02	0.02	0.56	0.01	0.37

	Anger	Contempt	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	0.59	0.00	0.10	0.04	0.00	0.02	0.14	0.10
Contempt	0.00	0.56	0.00	0.00	0.00	0.42	0.01	0.00
Disgust	0.23	0.05	0.54	0.01	0.00	0.15	0.00	0.01
Fear	0.00	0.01	0.00	0.44	0.00	0.34	0.00	0.22
Happiness	0.00	0.00	0.00	0.00	0.98	0.02	0.00	0.00
Neutral	0.00	0.00	0.00	0.00	0.04	0.96	0.00	0.00
Sadness	0.00	0.02	0.01	0.00	0.02	0.08	0.82	0.03
Surprise	0.00	0.00	0.17	0.02	0.02	0.37	0.00	0.58

Fig. 12. Normalized confusion tables for the emotional states. The table on the left represents the results obtained from [34]. The table on the right illustrates our results for mentioned emotional states.

therefore, a voice analysis would also be needed to capture data for those cases.

VII. Conclusion

Interestingly, as can be seen in the literature, several new methods and applications on HRI and their applications as a system of systems have been increasing in recent years. Due to their capability, strong connectivity, relatively smaller size, mobility, and interoperability, such methods provided a more extensive view on developing effective solutions not only in the HRI domain but also in any *human-in-the-loop scenarios* for Industry 4.0 studies. As each component of a system becomes more complex, the classification of the component as a system in itself is predominant, and hence the formulation of an HRI system as a system of systems is evident.

In the future, the application of the proposed system will be explored in the other aspects of the computational social systems, such as enhancing the social intelligence for the assistive robotic platform, user identification and recognition, and voice detection. In addition, more efforts will be performed to improve the computational efficiency of the proposed method. Although the work was evaluated with a limited number of users and aimed at a small social environment, we are planning to extend the work by introducing more users to interact with. Another objective for the future is the voice activation and user recognition by using the digital assistant device and the related web services, we believe this will lead us to explore more in the multimodality domain.

The paper presents an interactive heterogeneous robotic system with several components working together for closing the loop by detecting human emotions from facial images. The ultimate goal was designing a system of systems that fulfills user's requests by actively interacting with them, while other system components were gathering and sharing the data, and behaving as a whole to perform the given tasks. We have compared our results on particular focus points, such as computing the users' satisfaction and dissatisfaction levels by detected emotion states, within the literature and well-credited tools, such as the Microsoft Emotion API, to show the strength of our approach.

Our results show that the system works as designed and closes the loop in an HRI scenario. The emotion analysis performed by our algorithm clearly shows how macroemotions can be extracted to make better sense of the user's opinion. This becomes one of the key parts in closing the loop since this determines the branch of actions that the robot performs henceforth.

The results of the emotion transition prediction show that it performs very well when trained on a particular user. This is an expected scenario in an HRI domain since the use cases are designed for a limited set of people, e.g. in the case of a home-based assistive robot or heterogeneous inter-connected robotic platforms. Since we use transfer learning, the retraining of the network using user data can be done locally with limited processing power and resources on-board.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[2] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[3] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.

[4] A. R. Puthussery, K. P. Haradi, B. A. Erol, P. Benavidez, P. Rad, and M. Jamshidi, "A deep vision landmark framework for robot navigation," in *Proc. 12th Syst. Syst. Eng. Conf. (SoSE)*, Jun. 2017, pp. 1–6.

[5] P. Benavidez, M. Muppidi, P. Rad, J. J. Prevost, M. Jamshidi, and L. Brown, "Cloud-based realtime robotic visual slam," in *Proc. Annu. IEEE Syst. Conf. (SysCon)*, Apr. 2015, pp. 773–777.

[6] B. A. Erol, S. Vaishnav, J. D. Labrado, P. Benavidez, and M. Jamshidi, "Cloud-based control and vslam through cooperative mapping and localization," in *Proc. World Automat. Congr. (WAC)*, 2016, pp. 1–6.

[7] A. Karpathy. (2014). *What I Learned From Competing Against a Convnet on Imagenet*. [Online]. Available: http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 779–788.

[9] W. Liu *et al.*, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 21–37.

[10] M. A. Goodrich and D. R. Olsen, Jr., "Metrics for evaluating human-robot interactions," in *Proc. PERMIS*, 2003, p. 4.

[11] J. Scholtz, J. Young, J. L. Drury, and H. A. Yanco, "Evaluation of human-robot interaction awareness in search and rescue," in *Proc. IEEE Int. Conf. Robot. Automat.*, Apr./May 2004, pp. 2327–2332.

[12] A. Steinfeld *et al.*, "Common metrics for human-robot interaction," in *Proc. ACM SIGCHI/SIGART Conf. Hum.-Robot Interact.*, Salt Lake City, UT, USA, Mar. 2006, pp. 33–40.

[13] A. Steinfeld *et al.*, "Common metrics for human-robot interaction," in *Proc. 1st ACM SIGCHI/SIGART Conf. Hum.-Robot Interact.* New York, NY, USA: ACM, 2006, pp. 33–40.

[14] M. A. Goodrich and A. C. Schultz, "Human–robot interaction: A survey," *Found. Trends Hum.-Comput. Interact.*, vol. 1, no. 3, pp. 203–275, 2008.

[15] J. M. Beer, A. D. Fisk, and W. A. Rogers, "Toward a framework for levels of robot autonomy in human-robot interaction," *J. Hum.-Robot Interact.*, vol. 3, no. 2, pp. 74–99, 2014.

[16] P. Tsarouchi, S. Makris, and G. Chryssolouris, "Human–robot interaction review and challenges on task planning and programming," *Int. J. Comput. Integr. Manuf.*, vol. 29, no. 8, pp. 916–931, 2016.

[17] K. Dautenhahn, "Socially intelligent robots: Dimensions of human–robot interaction," *Phil. Trans. Roy. Soc. London B, Biol. Sci.*, vol. 362, no. 1480, pp. 679–704, 2007.

[18] Y. Song, L.-P. Morency, and R. Davis, "Multimodal human behavior analysis: Learning correlation and interaction across modalities," in *Proc. 14th ACM Int. Conf. Multimodal Interact.* New York, NY, USA: ACM, 2012, pp. 27–30.

[19] P. Ekman, W. Friesen, and J. Hager, "Facial action coding system: Research nexus," Netw. Res. Inf., Salt Lake City, UT, USA, Tech. Rep. 1, 2002.

[20] J. J.-J. Lien, T. Kanade, J. F. Cohn, and C.-C. Li, "Detection, tracking, and classification of action units in facial expression," *J. Robot. Auto. Syst.*, vol. 31, no. 3, pp. 131–146, 2000.

[21] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[22] M. Pardas, A. Bonafonte, and J. L. Landabaso, "Emotion recognition based on mpeg-4 facial animation parameters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, May 2002, pp. IV-3624–IV-3627.

[23] Y.-L. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 97–115, Feb. 2001.

[24] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.*, Apr. 1998, pp. 200–205.

[25] P. Michel and R. El Kaliouby, "Real time facial expression recognition in video using support vector machines," in *Proc. 5th Int. Conf. Multimodal Inter.* New York, NY, USA: ACM, 2003, pp. 258–264.

[26] B. Lwowski, P. Rad, and K.-K. R. Choo, "Geospatial event detection by grouping emotion contagion in social media," *IEEE Trans. Big Data*, to be published.

[27] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 2000.

[28] A. Jaimes and N. Sebe, "Multimodal human computer interaction: A survey," *Comput. Vis. Image Understand.*, vol. 108, no. 1–2, pp. 116–134, Oct./Nov. 2007.

[29] N. Sebe, I. Cohen, T. Gevers, and T. S. Huang, "Multimodal approaches for emotion recognition: A survey," vol. 5670, pp. 1–13, Jan. 2005. doi: 10.1117/12.600746.

[30] J. N. Bailenson *et al.*, "Real-time classification of evoked emotions using facial feature tracking and physiological responses," *Int. J. Hum.-Comput. Stud.*, vol. 66, no. 5, pp. 303–317, 2008.

[31] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Recognizing facial expression: Machine learning and application to spontaneous behavior," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2005, pp. 568–573.

[32] M. Rabiei and A. Gasparetto, "System and method for recognizing human emotion state based on analysis of speech and facial feature extraction; applications to human-robot interaction," in *Proc. 4th Int. Conf. Robot. Mechatron. (ICROM)*, Oct. 2016, pp. 266–271.

[33] D. R. Faria, M. Vieira, and F. C. Faria, "Towards the development of affective facial expression recognition for human-robot interaction," in *Proc. 10th Int. Conf. Pervas. Technol. Rel. Assistive Environ.* New York, NY, USA: ACM, 2017, pp. 300–304.

[34] *Microsoft Emotion Detector*. Accessed: Dec. 1, 2018. [Online]. Available: https://azure.microsoft.com/en-us/services/cognitive-services/emotion/

[35] M. Lapeyre, P. Rouanet, and P.-Y. Oudeyer, "The poppy humanoid robot: Leg design for biped locomotion," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Nov. 2013, pp. 349–356.

[36] B. A. Erol, A. Majumdar, J. Lwowski, P. Benavidez, P. Rad, and M. Jamshidi, "Improved deep neural network object tracking system for applications in home robotics," in *Computational Intelligence for Pattern Recognition*. Springer, 2018, pp. 369–395.

[37] Z. Kozhirbayev, B. A. Erol, A. Sharipbay, and M. Jamshidi, "Speaker recognition for robotic control via an IoT device," in *Proc. World Autom. Congr. (WAC)*, 2018, pp. 1–5.

[38] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. 1st Conf. Fairness, Accountability Transparency*, 2018, pp. 77–91.

[39] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.

[40] A. Dehghan, E. G. Ortiz, G. Shu, and S. Z. Masood, "DAGER: Deep age, gender and emotion recognition using convolutional neural network," Feb. 2017, *arXiv:1702.04280*. [Online]. Available: https://arxiv.org/abs/1702.04280

[41] G. Levi and T. Hassner, "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns," in *Proc. ACM Int. Conf. Multimodal Interact.* New York, NY, USA: ACM, 2015, pp. 503–510.

[42] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An all-in-one convolutional neural network for face analysis," Nov. 2016, *arXiv:1611.00851*. [Online]. Available: https://arxiv.org/abs/1611.00851

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

[44] F. A. Gers, J. A. Schmidhuber, and F. A. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Comput.*, vol. 12, no. 10, pp. 2451–2471, 1999.

[45] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with LSTM recurrent networks," *J. Mach. Learn. Res.*, vol. 3, no. 1, pp. 115–143, 2003.

Berat A. Erol (S'10) received the B.Sc. degree in mathematics from Kocaeli University, İzmit, Turkey, in 2007, and the M.Sc. degree in software engineering from St. Mary's University, San Antonio, TX, USA, in 2012. He is currently pursuing the Ph.D. degree with Autonomous Control Engineering (ACE) Laboratories, Department of Electrical and Computer Engineering, The University of Texas at San Antonio, San Antonio, TX, USA.

Since 2014, he has been a Graduate Research Assistant with the ACE Laboratories, Department of Electrical and Computer Engineering, University of Texas at San Antonio. His research activities at ACE Laboratories have been sponsored by the U.S. Department of Defense, Air Force Research Laboratory, and UTSA Lutcher Brown Endowed Chair. His current research interests include unmanned systems, HRIs, visual SLAM, machine learning, big data analytic, and Internet of Robotic Things (IoRT).

Mr. Erol is a member of Eta Kappa Nu (HKN) honor society.

Abhijit Majumdar (S'17) received the B.E. degree in electronics and communication engineering from the Shri Ramdeobaba College of Engineering and Management, Nagpur, India, in 2014. He is currently pursuing the M.S. degree in electrical engineering with the Autonomous Control Engineering Laboratories (ACE), Department of Electrical and Computer Engineering, The University of Texas at San Antonio, San Antonio, TX, USA.

He was a Graduate Research Assistant with the Autonomous Control Engineering (ACE) Laboratories, Department of Electrical and Computer Engineering, The University of Texas at San Antonio. From 2014 to 2016, he was with MG Automation Technologies, Nagpur, as a Research and Development Engineer. His research at the ACE Laboratories is supported by UTSA Lutcher Brown Endowed Chair. His current research interests include reinforcement learning and implementation of machine learning on robots and vision-based systems.

Patrick Benavidez (S'09–M'15) received the B.S., M.S., and Ph.D. degrees in electrical engineering from The University of Texas at San Antonio, San Antonio, TX, USA, in 2007, 2010, and 2015, respectively.

From 2007 to 2010, he was with Southwest Research Institute. He is currently an Assistant Professor of research and the Assistant Director of the Autonomous Control Engineering (ACE) Laboratories, The University of Texas at San Antonio. His doctoral research was supported primarily by Lutcher Brown Endowed Chair stipends, in part by Valero Research Excellence Awards, and in part by a competitive Graduate Student Research Scholarship. He has taken part in several grant writing efforts. He has mentored numerous students for Ph.D. dissertation, master's thesis, projects, and undergraduate capstone projects. He is an Investigator on a U.S. $5 million AFRL Center of Excellence in Autonomy Grant, and a Principal Investigator on a U.S. $130 thousand contract to develop a machine learning system for a security application. His current research interests include machine learning, communication systems, control systems, robotics, cyber-physical systems, and systems of systems.

Dr. Benavidez received several awards for volunteering and outreach during his Ph.D. studies, including Most Exceptional Graduate Student Award from the UTSA College of Engineering and a University Life Award for Most Outstanding Graduate Student in the College of Engineering from the UTSA Student Government.

Paul Rad is currently an Associate Professor with the Department of Information Systems and Cyber Security and the Co-Founder of the Open Cloud Institutes (OCI), The University of Texas at San Antonio (UTSA), San Antonio, TX, USA. He has a recognized record of entrepreneurship, business technology, and academic contributions in cloud computing, cyber security, and artificial intelligence (AI). He holds advisory board seats at several cutting-edge startups on cloud computing and AI. He has conducted research on several important topics such as knowledge representation, vision, and natural language understanding. His research has been extensively funded by the NSF, DoD, State of Texas, and enterprises such as Facebook, Intel, Dell/EMC, Cisco, USAA, and Rackspace. He holds 16 U.S. patents. His current research interests include computational intelligence (CI), resilient algorithm, and safe decision-making for autonomous systems.

Mr. Rad was named the High-Performance Cloud Group Chair at Cloud Advisory Council in 2014. He and his collaborators received the first U.S. $10 million National Science Foundation (NSF) cloud grant in 2015.

Kim-Kwang Raymond Choo (SM'15) received the Ph.D. degree in information security from the Queensland University of Technology, Brisbane, QLD, Australia, in 2006.

He is currently the Cloud Technology Endowed Associate Professor with The University of Texas at San Antonio (UTSA), San Antonio, TX, USA.

Dr. Choo is a fellow of the Australian Computer Society. He served as the Co-Chair of IEEE Multimedia Communications Technical Committee's Digital Rights Management for Multimedia Interest Group. In 2016, he was named the Cybersecurity Educator of the Year—APAC (Cybersecurity Excellence Awards are produced in cooperation with the Information Security Community on LinkedIn). He and his team won the 2015 Digital Forensics Research Challenge organized by Germany's University of Erlangen-Nuremberg. He is the recipient of the 2018 UTSA College of Business Col. Jean Piccione and Lt. Col. Philip Piccione Endowed Research Award for Tenured Faculty; the Outstanding Associate Editor of 2018 for IEEE ACCESS; the British Computer Society's 2019 Wilkes Award Runner-Up Award; the 2019 EURASIP Journal on Wireless Communications and Networking (JWCN) Best Paper Award; the Korea Information Processing Society's Journal of Information Processing Systems (JIPS) Survey Paper Award (Gold) in 2019, the IEEE TrustCom 2018 Best Paper Award; the ESORICS 2015 Best Research Paper Award; the 2014 Highly Commended Award by the Australia New Zealand Policing Advisory Agency; the Fulbright Scholarship in 2009; the 2008 Australia Day Achievement Medallion; and British Computer Society's Wilkes Award in 2008.

Mo Jamshidi (S'66–M'67–SM'74–F'89–LF'09) received the B.S. degree in electrical engineering (EE) from the Oregon State University, Corvallis, OR, USA, in 1967, and the M.S. and Ph.D. degrees in EE from the University of Illinois at Urbana–Champaign, Champaign, IL, USA, in 1969 and 1971, respectively, the Honorary Doctorate degrees from the University of Waterloo, Waterloo, ON, Canada, in 2004; the Technical University of Crete, Chania, Greece, in 2004; and Odlar Yourdu University, Baku, Azerbaijan, in 1999.

He has been an Advisor to NASA (including 1st MARS Mission), HQR, USAF Research Laboratory, Kirkland Air Force Base, USDOE, Oak Ridge, TN, USA, and EC/EU, Brussels, Belgium. He is currently the Lutcher Brown Endowed Chair and a Distinguished Professor with The University of Texas at San Antonio, San Antonio, TX, USA. He is also an Honorary or Visiting Professor with Deakin University, Waurn Ponds, VIC, Australia; Birmingham University, Birmingham, U.K.; Obuda University, Budapest, Hungary; Loughbrough University, Loughbrough, U.K.; East China Normal University, Shanghai, China; Nanjing University, The Nanjing, China, Xian University, Xian, China. He has advised more than 70 M.S. and 65 Ph.D. students. He is currently involved in research on system of systems engineering with emphasis on cloud computing, robotics, UAVs, biological and sustainable energy systems, including smart grids, machine learning and big data analytic. He has authored or coauthored more than 800 technical publications, including 74 books (11 text books), research volumes, and edited volumes in English and 7 foreign languages.

Dr. Jamshidi has been a member of the University of the Texas System Chancelors Council since 2011. He is an A. Fellow of AIAA and TWAS, a fellow of AAAS, ASME, and NYAS, and a member of HAE. He was a recipient of numerous honors and awards, including the IEEE Centennial Medal, the IEEE Millennium Awards, and the IEEEs Norbert Weiner Research Achievement Award, the 2014 IEEE-USA Systems Engineering Career Award, and so on. He is the Founding Editor or Co-Founding Editor or Editor-in-Chief of five journals including *IEEE Control Systems Magazine* and the IEEE SYSTEMS JOURNAL. He has over 10 500 Google Scholar citations.