IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 23, 2022, accepted June 3, 2022, date of publication June 13, 2022, date of current version June 16, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3182469

Multimodal Engagement Prediction in Multiperson Human–Robot Interaction

AHMED A. ABDELRAHMAN, DOMINYKAS STRAZDAS, (Graduate Student Member, IEEE), ALY KHALIFA, JAN HINTZ, THORSTEN HEMPEL, AND AYOUB AL-HAMADI

Neuro-Information Technology, Otto-von Guericke University Magdeburg, 39106 Magdeburg, Germany

Corresponding author: Ahmed A. Abdelrahman (ahmed.abdelrahman@ovgu.de)

This work supported by the Federal Ministry of Education and Research of Germany (BMBF) within the Zwanzig20 Alliance 3Dsensation under Grant RoboAssist 03ZZ0448L and Grant Robo-Laboratory 03ZZ04X02B.

ABSTRACT The ability to measure the engagement level of humans interacting with robots paves the way towards intuitive and safe human-robot interaction. Recent approaches achieve reasonable progress in predicting human engagement in physically situated environments. However, engagement estimation is still a challenging problem especially in an open-world environment due to the difficulty of creating and monitoring a variety of human social cues in real-time. Furthermore, the interactions may involve a group of subjects interacting simultaneously with the robot, which increases the prediction complexity. In this paper, we design a real-time engagement estimation system for humans interacting with robots with generalization capability. We propose to estimate engagement using a three-stage approach based on a combination of learning-based and rule-based approaches. Firstly, state-of-the-art deep learning methods are used to extract engagement features from input frames. Then, a simple neural network is used to estimate the focus of attention score by incorporating gaze and head pose features and assigning this score to all subjects in the scene using a face recognition algorithm. Finally, a rule-based classification approach is used to predict the engagement state of the subject to initiate/terminate the interaction with the robot. To effectively evaluate our system, we access our approach for each phase separately. Additionally, we use an online evaluation study in which subjects are allowed to interact freely with an industrial robot. Our model achieves an average of 96%, 90%, and 93% precision, recall, and F-score respectively.

INDEX TERMS Human–robot interaction, engagement, disengagement, cobots, rule-based approach, deep learning, face recognition, gaze estimation, head pose, body pose.

I. INTRODUCTION

Human-Robot Interaction (HRI) [1]–[3] arises to tackle and enhance the meaningful interaction between humans and robots while preserving the safety of human interaction partners. These interactions include three stages of collaboration: coexistence, cooperation and responsive collaboration. Consequently, a new type of industrial robot called collaborative robot (cobot) is developed with a high level of safety to join the industry along with humans.

Using cobots aside with Artificial Intelligence opens the way for the development and creation of natural and intuitive HRI concepts. In this paper, we develop a concept that combines different features including head pose, gaze, posture, speech, and gesture to allow a multimodal and restriction-free HRI. The interactions between humans and robots are divided into four different phases: Firstly, the robot detects a potentially interested user for engagement; Secondly, the robot identifies the kind of interaction intention from the person; thirdly, the robot executes the task or gives a response to the person; Finally, human disengages from the robot after receiving the response, or after task execution. The study of engagement and disengagement phases are crucial for maintaining sustainable spontaneous interactions between humans and robots [4].

Engagement is a complex process that can be analyzed in terms of four discrete stages [5] include the intention to engage, engagement, disengagement, and re-engagement.

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

The re-engagement phase only occurs when the disengagement is not complete which is very difficult to determine with certainty. The estimation of the engagement state requires the challenging tracking of a variety of social cues during interactions. The most common signals used in literature are gaze [6]–[10], head motion [9]–[11], body posture [12], and distance [12].

Recently, machine learning-based algorithms have been greatly used to classify engagement including logistic regression [11], boosted decision tree [13], maximum entropy model [14], Support Vector Machine (SVM) [15], and neural networks [16]. However, they largely depend on annotating new datasets with engagement states which is difficult, expensive, and requires trained annotators [11]. On the other hand, the rule-based approaches had a competitive performance compared to machine learning classifiers trained on a labeled corpus [17].

classify engagement states

In this paper, we develop a new approach that utilizes simple rule-based policies to predict the human engagement state interacting with a cobot seen in Fig. 1. Our model depends on data acquired from sensors that are readily available on different robot systems. Our system consists of three stages including feature extraction, feature processing, and engagement classification stages. Firstly, features were extracted using state-of-the-art deep learning techniques including gaze, head pose, body posture, and face identification (ID). Secondly, we use a feedforward neural network to estimate the focus of attention score of all subjects in the scene using head pose and gaze. Further, we adopt a face recognition algorithm for person identification to temporally monitor subjects in a multi-user interaction. When a potential user is identified, the corresponding face ID is matched with the closest detected body using a rule-based policy. Finally, a simple rule-based classification approach is used to predict engagement and disengagement to initiate/terminate the interaction of the subject with the robot. This way, we introduce a contactless human-machine interaction approach by enabling the interaction only when intended, leading to a safer HRI.

Our contributions can be summarized as follows:

- We design a novel model to predict human engagement state during interaction with robots. Further, our model has a generalization capability that can work with different robot systems because it depends on simple rule-based policies which do not require costly annotations.
- We introduce a person identification method using a face recognition algorithm to keep track of the person's identity during the whole interaction to classify disengagement and re-engagement classes.
- We evaluate our model using an online user study with subjects interacting freely with a cobot. Further, we provide the benchmarks for the methods adopted in each stage of our model separately as well as a subject assessment questionnaire.

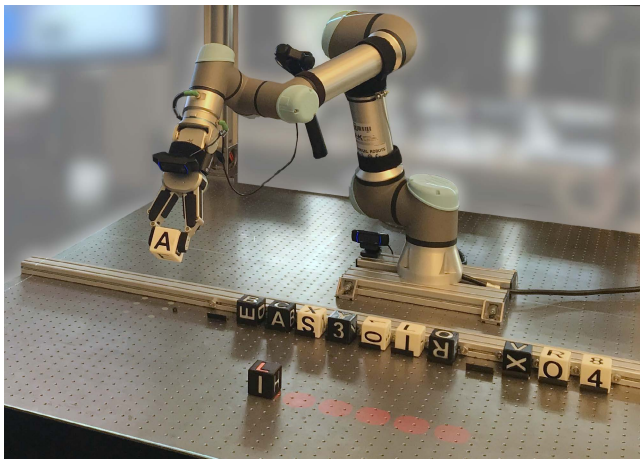


FIGURE 1. Cobot: UR5e industrial robot and cubes used in the experiments.

II. RELATED WORK

In this section, we will outline the related work regarding industrial human-robot interaction, engagement, and disengagement techniques in HRI.

A. INDUSTRIAL HUMAN-ROBOT INTERACTION

The number of industrial robots joining the industry is increasing rapidly in the last decades. They are used in various applications include welding, disassembly [18], pick and place for printed circuit boards, and transportation [19]. As a special type of industrial robot, cobots are lightweight robots equipped with different safety features and designed for direct physical interaction with humans [20]. Ensuring human safety is one of the most crucial aspects while interacting with robots. Physical safety is one of the most important ways of human safety in HRI [1]. It targets to maintain no unwanted or unintentional direct physical contact between humans and robots. Recently, researchers developed a lot of methods and techniques to ensure human safety while interacting with robots [21]. Human engagement estimation is one of the methods that help ensure safety in HRI by enabling interaction only when intended. Sidner *et al.* [22] define engagement as ''the process by which individuals in an interaction start, maintain and end their perceived connection to one another''. A lot of methods and techniques are developed to predict the engagement state of humans interacting with robots.

B. ENGAGEMENT TECHNIQUES IN HRI

Engagement occurs when humans and robots start and maintain their perceived interaction. In general, the engagement starts based on detecting an intention from a human to engage with the robot [23]. In [15] and [17], the authors studied human engagement by focusing on the angle of the user engagement state prediction. Castellano *et al.* [15] showed that the integration of game and social features can lead to improved prediction performance. Foster *et al.* [17] estimate the engagement state of customers for a robot bartender

using visual and audio data. They proposed two different methods for predicting engagement, including a rule-based method and trained classifiers. Further, they evaluated their methods using an online evaluation scenario with real subjects along with the offline evaluation. They concluded that the simple, rule-based classifier achieves competitive performance compared with multiple trained classifiers in all the experiments. Vaufreydaz *et al.* [12] demonstrated that multimodal sensor data (including 32 features from multiple sensors) provides better performance than using only spatial features. Furthermore, they illustrated that seven selected features only are adequate to provide good performance for engagement detection. Li *et al.* [24] developed an active method where the robot interacts naturally with multiple persons. They reported that their method is capable of selecting a person as an addressee based on the perception of human visual cues. In [14], the authors created models for automatically identifying human intentions with low false-positive rates and before the engagement (3-4 seconds earlier). They used audiovisual modalities to calculate human intentions for initiating the interaction with a robot. Schuller *et al.* [25] offered an audiovisual approach for the identification of spontaneous engagement in human conversations. They demonstrated that balancing the training sets leads to significantly better results and also found that the combination of the audio and visual modality is better than using the single modalities, and is still real-time capable. Richter *et al.* [26] studied the impact of addressee identification on the robot's performance to classify utterances directed to it from an interaction between humans. They evaluated methods to addressee identification by utilizing different types of visual cues. Reference [14] presented a computational model for managing engagement with a situated agent in multiparty, open-world settings.

C. DISENGAGEMENT TECHNIQUES IN HRI

Engagement breakdown occurs when humans are about to terminate their interactions with the robot. Predicting disengagement in open-world settings is a challenging problem as it involves the processing of multiple signals, including head pose, eye gaze, posture, and gestures. Reference [13] proposed a self-supervised approach to forecast disengagement. They construct forecasting models that do not require manual annotations by deploying the robot NAO with a baseline forecasting model in an open space for five days. Furthermore, they trained logistic regression and boosted decision tree models using the data collected from 158 users interacting with the robot and compared it with the forecasting model. They conclude that the forecasting model can predict disengagement while maintaining a low false-positive rate.

Leite *et al.* [9] proposed three different SVM-based models for predicting disengagement of humans when interacting with social robots: a model trained with a dataset containing one subject engaging with the robot, a model trained with a dataset containing a group of three subjects engaging with the robot, a model trained with combined instances from the two datasets. The authors conclude that the model trained only with data from one engagement user might not be appropriate for group engagements. However, a model trained only with group data performs well when tested with a single user. Furthermore, they showed that the mixed model performed better than the two models [9].

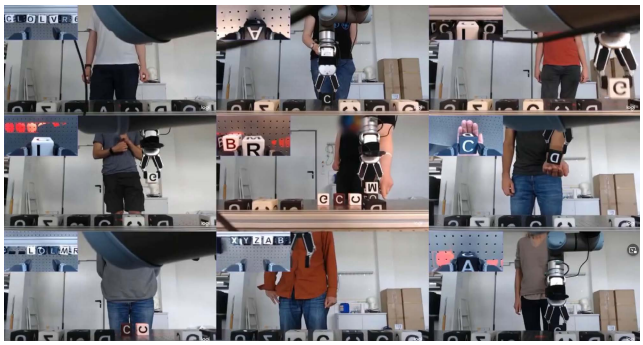


FIGURE 2. Previous Wizard-of-Oz study [27]. A video summary can be found here: https://youtu.be/JL409R7YQa0.

Finally, Ben *et al.* [11], [16], [28] presented a new dataset called UE-HRI for spontaneous interactions between humans and a Pepper robot. In [16], deep learning techniques are utilized (deep and recurrent neural networks) to predict disengagement in real-time, achieving a reasonable prediction accuracy of 78%. In addition, they further investigate the prediction of disengagement in [11] by investigating the time interval over which a user's disengagement can be detected using supervised classifiers. They concluded that the used models were generally successful in predicting the disengagement up to 10 seconds before the user leaves the interaction.

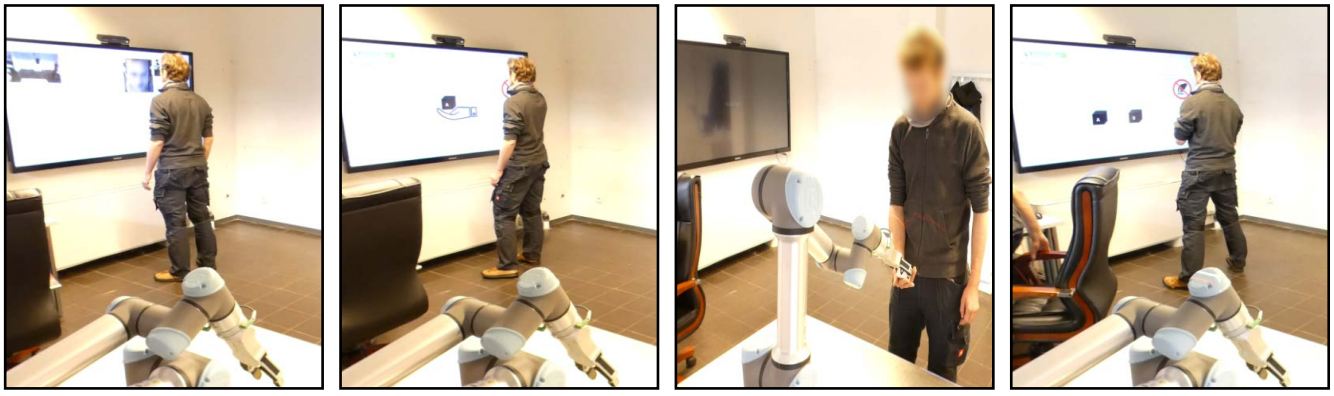
III. COBOT SYSTEM

In our earlier Wizard-of-Oz (WoZ) study [27], we designed a mockup of a Robot System Assistant and carried out a study with 36 subjects. The subjects were permitted to use different features like gestures, speech, mimics, and gaze without any limitations to communicate with a cobot and execute different tasks like cube stacking. The subjects think that they are interacting with an artificial system, while in reality, we were controlling the cobot from a hidden room, based on the subject's instructions. Figure 2 shows a few interactions between subjects and the cobot from our WoZ study.

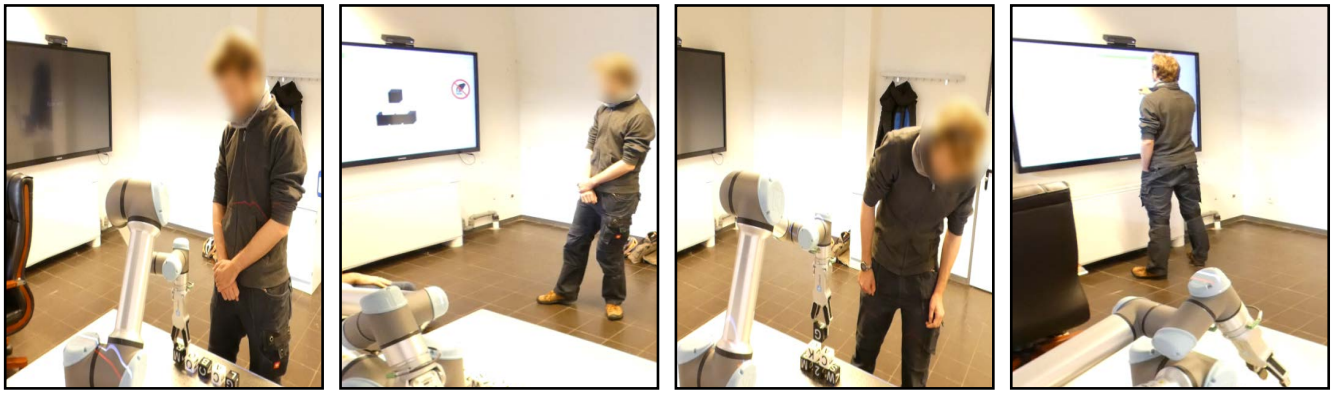
In this paper, we conduct an online user study in the same way as in the WoZ study in order to evaluate our engagement and disengagement model in an open-world environment.

A. EXPERIMENTAL SETUP

Fig. 4 shows the general hardware resources used in the experimental design of the interaction. The hardware resources include two different workstations WS1 (cobot) and WS2 each of them having a different design and purpose.



(a) Subject registration on WS2 (b) Read task 1 instructions on WS2 (c) Execute task 1 with WS1 (d) Read task 2 instructions on WS2



(e) Execute task 2 with WS1 (f) Read task 3 instructions on WS2 (g) Execute task 3 with WS1 (h) Answer questionnaires on WS2

FIGURE 3. Typical human-robot interactions scenario with three phases including registration, interaction, and questionnaires.

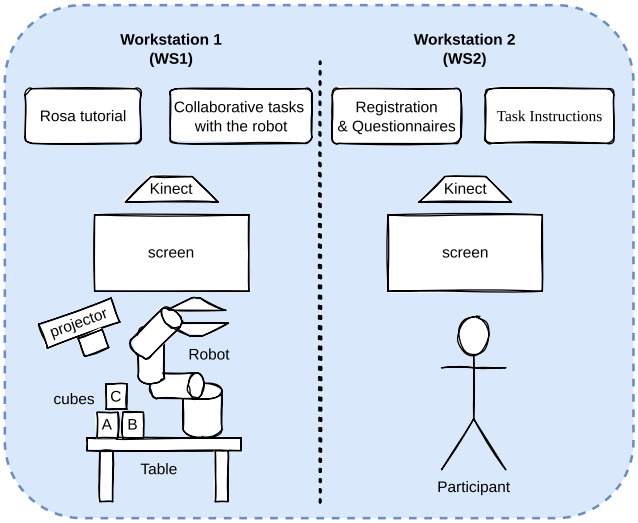


FIGURE 4. A schematic overview of the workstations (WS1 and WS2).

1) WS1

is responsible for the main interaction between the user and the cobot. It consists of an *UR5e* industrial robot with a *RG6* gripper placed on a metal table in front of a TV screen. A time-of-flight (ToF) Kinect V2 camera is placed above the screen so that the user is seen fully in front of the table. On the table, there are black and white letter cubes, which can be picked up by the robot.

2) WS2

is mainly responsible for registering the subject's facial features in the system which is crucial for the engagement estimation. Further, It is used for the questionnaires at the end of the experiments. It consists of a touch-capable smart screen with built-in speakers and another Kinect camera equipped with a microphone.

B. SCENARIO

The scenario of the experiments is divided into 3 phases as shown in Fig. 3 including registration, interaction, and the questionnaires.

1) REGISTRATION

The subject uses WS2 for registration as shown in Fig. 3(a). WS2 has an interface based on the ROS-QT visualization environment. The subject uses this interface to register his name, preferred hand, and face features to the system. In order to store all facial features, we ask the subject to look to the front and then once to the right and once to the left through

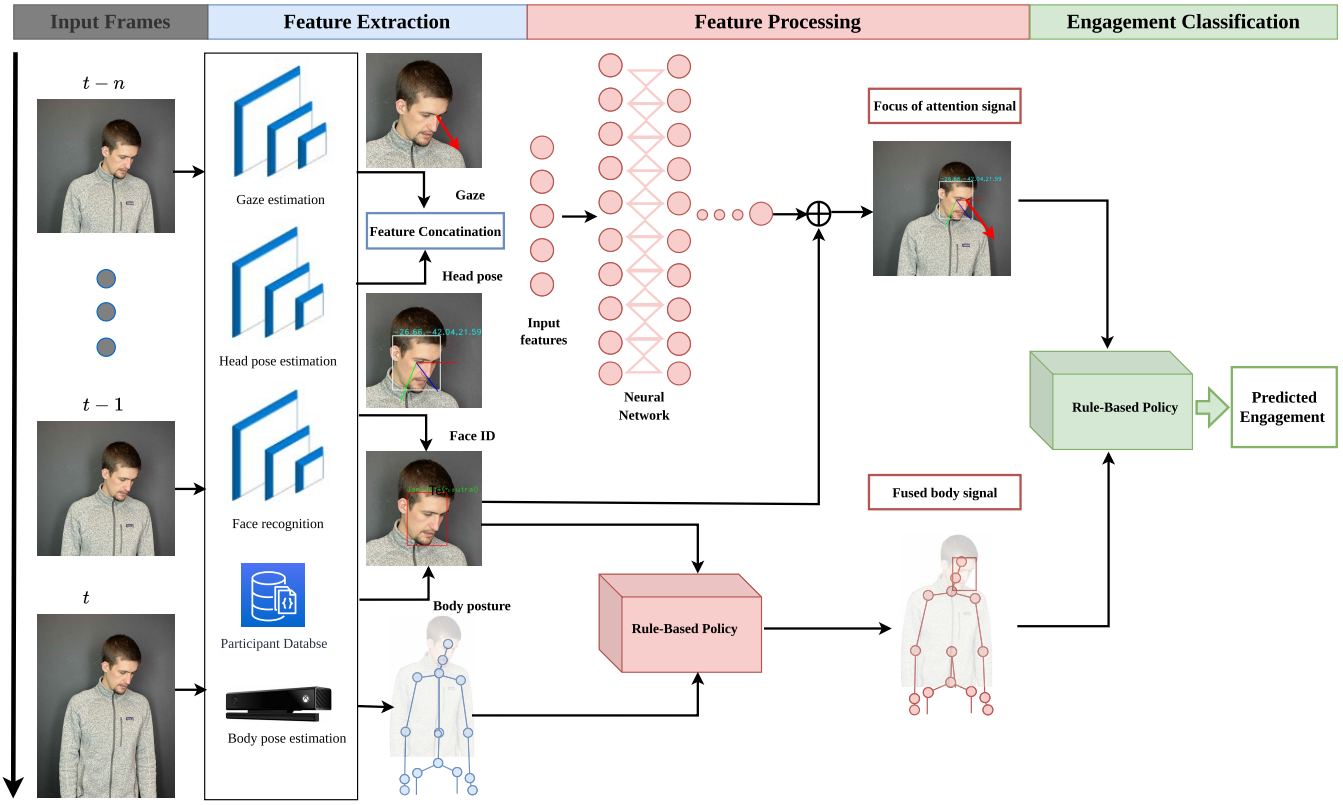


FIGURE 5. Overall architecture of our multi-modal engagement prediction model including feature extraction, feature processing, and engagement classification phases.

the interface. Based on the stored face features, we assign a face ID to the subject that will be used during the whole experiment.

2) INTERACTION

The subject switches to WS1 to start doing three different collaborative tasks with the cobot including:

1) Have Cobot give you a block as in Fig. 3(c).
2) Spell a specific word with the alternating color of blocks as in Fig. 3(e).
3) Build a 3-2-1-Pyramid with black-white-black layers as in Fig. 3(g).

We add one more step to this phase that the subject switches to WS2 to read the instruction for the next task and then returns to WS1 to execute the task. This step will add more challenges and complexity to the deployed engagement model.

3) QUESTIONNAIRES

Finally, the subject switches to WS2 to answer questionnaires as in Fig. 3(h). Additionally, a module-specific questionnaire was taken to allow the subjects to evaluate each module separately after finishing the experiment.

IV. APPROACH

We propose a new model that utilizes simple rule-based policies to predict the human engagement state of humans interacting with a cobot. Our model is easy to use approach as it employs visual data that is available on various robot platforms. It can track the engagement state of humans starting from willing to engage until disengaging. Further, our model can predict the re-engagement state based on the subject identity features obtained from a face recognition algorithm. Our model consists of three stages include feature extraction, feature processing, and engagement classification:

1) **Stage 1**: we extract crucial features for classifying engagement including human head pose, gaze, and face ID using state-of-the-art convolution neural networks. Further, we use Kinect v2 to extract the body pose feature.
2) **Stage 2**: we output the most important signals for the engagement including the focus of attention and fused body signals using a feed-forward neural network.
3) **Stage 3**: we classify the engagement state of subjects using a simple rule-based policy based on the output of stage number two.

Fig. 5 shows the overall architecture of our multi-modal three-stage engagement and disengagement prediction model.

A. FEATURE EXTRACTION

Various cues could represent the subject's engagement intentions, which could be used to model engagement and

TABLE 1. Extracted feature streams.

Stream	Feature	Description
Gaze	Gaze	2 features [yaw, pitch] (in degrees)
Head	Head	3 features [yaw, pitch, roll] (in degrees)
Face	Face ID	1 feature $\in \mathbb{Z}_{>0}$
Body	Body Joints	25 features [x,y,z]

disengagement. Using more features to estimate engagement may increase the prediction accuracy. However, it will increase the computational cost, which will harm the overall model performance. Consequently, we extract the most crucial features to characterize engagement as mentioned in most of the literature including gaze, head pose, body posture, and face ID as shown in Table 1. To extract robust features, we apply state-of-the-art deep learning methods to the raw signals obtained from the resources.

1) GAZE

Gaze is the most crucial cue used in predicting engagement and disengagement, as it defines the person's current visual focus of attention. The direction of gaze defines the initial intention of humans to engage with the robot. We propose to use a robust gaze estimation method designed by Petr *et al.* [29] which they extract gaze using a convolution neural network (CNN). As seen in the Table 1, we estimate 3D gaze direction as spherical coordinates (pitch, yaw) relative to the camera view.

2) HEAD POSE

Head pose is a good approximation of the person's focus of attention, especially when the gaze is missed or inaccurate. The direction of the head pose combined with the gaze can deduce the interaction willingness of the person to start/terminate the interaction with the robot. A lot of methods and techniques were proposed to estimate 3D face poses from images. We propose to use a robust method developed by albiero *et al.* [30] which is a real-time, six degree of freedom (6DoF) head pose estimation. Then, we convert the 6DoF estimated by the model to the head pose Euler angles (pitch, roll, yaw) as seen in Tab. 1.

3) FACE ID

Person identification is crucial, especially when dealing with multi-subject HRI. We adopt a face recognition framework based on three stages including face detection, face alignment, and face identification. Firstly, we utilize RetinaFace [31] to detect and localize the faces in the input frames, which is one of the most accurate face detection methods. Furthermore, we use MobileNet-0.25 [32] as a backbone to achieve real-time HRI. Secondly, the bounding boxes with high confidence are aligned by a practical facial landmark detector [33]. The aligned faces are cropped to a size of 112 × 112 to be more consistence with the next stage. Finally, the face features are extracted from the cropped faces by means of the ArcFace model [34] and outputs the corresponding feature embedding vector of 512 features. We compare the embedding vector against the gallery embedding to predict the face ID as shown in Tab. 1.

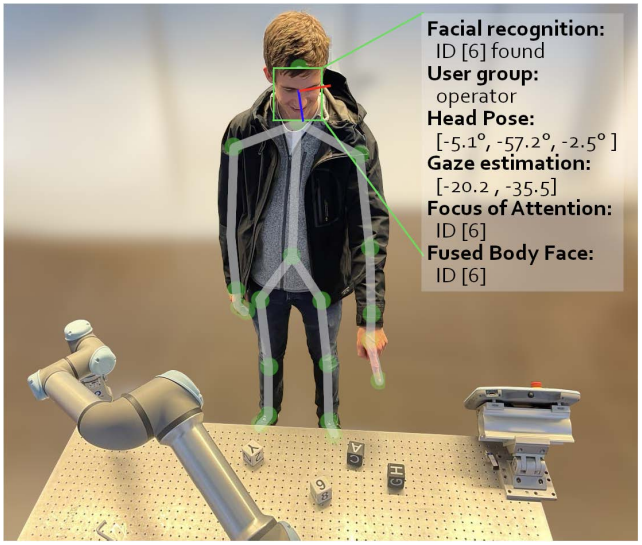


FIGURE 6. Crucial features extracted during the experiments used to model engagement including gaze, head pose, body posture, and face ID.

To the best of our knowledge, all related works that estimated engagement and disengagement do not consider the face identification feature in their techniques.

4) BODY POSTURE

The body posture provides a good indication of the person's engagement degree, especially when the face is missed. Body posture is estimated with the Kinect v2 using the native Kinect for Windows SDK. The resulting skeleton contains 25 points mapped in the 3D coordinate system relative to Kinect. Each joint can have a tracking state: "tracked", "inferred" or "not tracked", thus determining the overall quality of the tracked skeleton.

Fig. 6 presents the four features extracted from streams in order to predict engagement and disengagement.

B. FEATURE PROCESSING

Using the extracted features, we generate two important binary signals to characterize engagement. These signals include the focus of attention (FA) and fused body face (BF), which are crucial for engagement estimation. Our model uses these signals aside with rule-based policy to make engagement decisions.

1) FOCUS OF ATTENTION SIGNAL

In every frame, a simple model is used to deduce whether the attention of each subject in the scene is pointed to the (WS) robot or not. This inference is currently based on a simple neural network model trained using a manually labeled

Algorithm 1 Skeleton S and Face $F\{x, y, ID\}$ Fusion

Faces ←All detected Faces
Bodies ←All tracked Kinect Skeletons
for all S in *Bodies* **do**
    x_S, y_S ←Transform$_{2D}$(S.*Head* $\{x, y, z\}$)
    $d \leftarrow \infty$
    for all F in *Faces* **do**
        $\delta_F \leftarrow \sqrt{(x_s - F.x)^2 + (y_s - F.y)^2}$
        if $\delta_F < d$ **and** $\delta_F < 30$ pixels **then**
            $d \leftarrow \delta_F$
            S.ID $\leftarrow F$.ID
        end if
    end for
    if S.ID $\neq$ *null* **then**
        remove $F(S.\text{ID})$ **from** *Faces*
        Users.add(S)
    end if
end for

Algorithm 2 Engagement

Users ← All fused bodies
A ← Focus of Attention
for all Workstations W **do**
    if W.*current_user* $\neq$ *null* **then** continue
    end if
    for all Fused Bodies B in Users **do**
        if $B.ID == A.ID$ **then**
            if is_registered(B) **then**
                W.*current_user* ← B
                welcome(B)
            else
                send_to_registration(B)
            end if
        end if
    end for
end for

dataset. We concatenate the extracted 3D gaze direction and 3D head pose angles to form a 5-dimensional feature vector. We build a simple classification neural network containing 6 fully connected layers followed by a ReLU activation layer. The network takes the gaze and head pose vector as an input and outputs a binary attention signal. We create an attention score counter for each subject based on the face ID feature. At every frame, the attention score counter is incremented if the face of the subject is frontal and the gaze is directed to the robot and decremented vice-verse. If the attention score reached a specific threshold we publish the subject ID as a ROS message "FA" reporting that a specific subject is focusing its attention on the robot.

2) FUSED BODY SIGNAL

To prevent an accidental mix-up of the detected skeletons and to prohibit (accidental/mischievous) input from a skeleton, not engaged with the system, we only consider users who have a valid skeleton and a valid registered face ID. We refer to the combination of the two as "fused body". Consequently, face ID to skeleton matching plays a crucial role in user management, safety and is necessary in order to maintain an efficient interaction between one operator and the system. It also plays an important role in multi-user scenarios where each input must be matched to a corresponding user.

The algorithm 1 takes a valid and tracked skeleton and transforms its head coordinates into 2D camera space. For the corresponding head coordinates, the closest face is found and checked for deviation (<30 pixels). The face ID is then matched to the body and the face is then removed from the face array to prevent double IDs. Only a body with an ID, referred as fused body, is further considered for interaction and are considered for input. For each Frame, the fused body is checked for integrity. If the matched face or the skeleton is missing for more than 3 frames, the fused body is removed from the user array. Since the face ID is the primal identification feature and is by definition constant, it is only necessary to track this ID for any interaction.

C. ENGAGEMENT CLASSIFICATION

We model the problem of engagement and disengagement as a binary classification. Using the focus of attention and fused body signals, we create a rule-based approach to predict engagement in each workstation. Each workstation monitors its current active engaged subject and the corresponding ID. The system is designed according to two constraints including:

- The system must be free, i.e., no one is already engaged with it.
- The user must be registered and, if necessary, have an appropriate clearance for the workstation.

1) ENGAGEMENT

The engagement process, as explained in algorithm 2 happens automatically when a subject starts interacting with the system and thus accumulates a focus of attention and fused body. If multiple subjects are engaging at the same time, the one with the highest focus of attention score gets chosen. The system reacts to the engagement process by turning on the monitor, turning the robot towards the subject, and greets the subject with the name, chosen during the registration process. If the person engaging with the system is unknown, the system guides him to WS2 for the registration process.

2) DISENGAGEMENT

The examination for disengagement, the process explained in algorithm 3, starts after a subject engages with a system and is repeated every frame. For the interaction to take place, a valid fused body and positive focus of attention scores are necessary. If either of them is missing, a reminder timer gets started. If either of the features returns, then the timer is stopped and the subject notices no difference in

Algorithm 3 Disengagement

```
Users ← All fused bodies
A ← Focus of Attention
for all Workstations W do
    B ← W.current_user
    if B == null then continue
    end if
    if A = nullorB ∉ Users then
        start disengagement timer
        display message: "Are you still there?"
    end if
    if disengagement timer timeout then
        W.current_user ← null
    end if
    if B.ID == A.ID then
        stop disengagement timer
    end if
end for
```

interaction. In the case of a timeout, a message "Are you still there?" is displayed, asking the subject to return to the scene or to become attentive again. During the message display, a logout timer is started, after which the subject ID is removed as active from the corresponding workstation if the focus of attention or the fused body-face are still missing.

After the disengagement, the robot and the monitors go to an idle state or the next attentive subject fulfilling the prerequisites becomes engaged with the system.

V. RESULTS

Our engagement prediction model is a three-stage approach including feature extraction, feature processing, and engagement classification. To effectively assess our model, we access our approach for each phase separately focusing on the methods and techniques used during this stage.

A. FEATURE EXTRACTION

We present the evaluation to the CNNs used to extract four engagement features used in our model.

1) GAZE

We use a CNN that consists of a ResNet-18 as a backbone followed by two fully connected layers for outputting gaze direction. We utilize pinball loss as a network loss function that estimates the gaze direction and error bounds together, which improves the gaze performance. Further, we train our model using the gaze360 dataset which is collected under unconstrained settings, to improve our model generalization capabilities. We follow the same evaluation criteria as in [29] by dividing the dataset into train-val-test sets. We train the network for 100 epochs with a batch size of 80 and a constant learning rate of 0.0001. Our gaze estimation model achieves a performance of 15.3 ° mean angular error on the gaze360 dataset.

2) HEAD POSE

We utilize a CNN that consists of a ResNet-18 as a backbone and exploit a stochastic gradient descent with a mini-batch of two images. Further, we train our model using the WIDER FACE dataset [35], which contains image samples with high diversity in scale, pose, and occlusion. We train the model for 35 epochs with a batch size of 512 and a variable learning rate starts from 0.001 and decreased by a factor of 10 if the performance does not increase over 3 epochs. Our head pose estimation model achieves a performance of 3.913 mean absolute error on the AFLW2000-3D dataset which is a challenging dataset with a diversity of illumination, head pose, and facial expression.

3) FACE ID

We use a deep CNN for recognition feature extraction. To achieve real-time inferring, we utilize MobileFaceNet [36] as a backbone learned by ArcFace [34] loss function. MobileFaceNet can infer face feature embedding within 2.4ms with a model size of 112MB. Further, we train our model using the MS-Celeb-1M dataset [37] which contains about 100k identities with 10 million images. The model was trained for 30 epochs with a batch size of 512 and a variable learning rate starting from 0.001 and decreased by a factor of 10 at 100K, and 160K iterations. Our face recognition model achieves a performance of 95.45% and 92.07% on CPLFW and CALFW datasets respectively. CPLFW and CALFW datasets show higher pose and age variations with the same identities as the LFW dataset.

4) BODY POSTURE

We use the native Kinect2 for Windows SDK to estimate the 25 points for body posture. Further, we removed any skeletons containing more than eight inferred points for a short period of time to deal with the problem of Kinect falsely detecting skeletons in inanimate objects. This is due to the fact that the false detections, while having a stable torso, usually do not have a stable limb detection. These false limbs are then estimated as vigorously shaking around, while having the "inferred" status. To further reduce a possible false-positive body detection, a rage for expected body position values, or basically an "interaction zone" of $2m \times 2m \times 2m$ was set up. No chairs or doors or other inanimate objects were considered as bodies after this implementation, while the real person detection rate was not reduced.

B. FOCUS OF ATTENTION

Furthermore, we evaluate the focus of attention neural network as It is an important stage in our model. We use binary labels for modeling the FA which includes two classes: looking toward the robot and looking somewhere else. To train and evaluate our network, we used two public datasets with gaze and head pose annotations: Gaze360 [29] and BIWI [38]. Based on the range of head

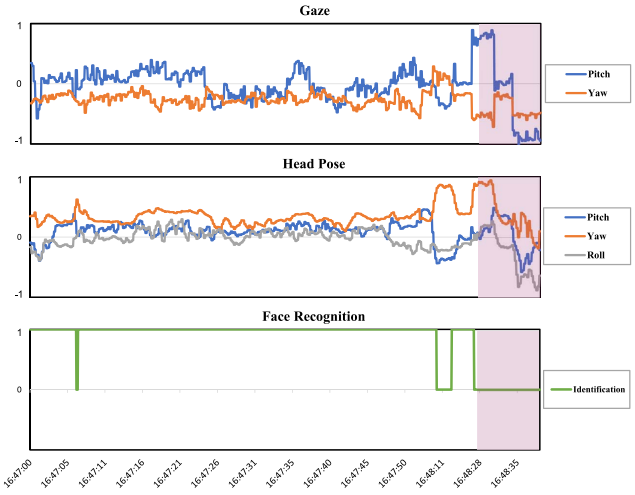


FIGURE 7. Example of multimodal features recorded using robot time-stamps. Pink color presents the end of the interaction (engagement).

pose angles and gaze direction, we manually annotated the FA for 50400 images in two classes: looking towards the robot (35367 images) and looking somewhere else (15033 images). The annotated dataset distribution is unbalanced and biased toward the first class. To avoid this problem, we apply oversampling technique by increasing the number of instances of the second class. We split the annotated data into 2 parts: 80% for training and 20% for testing. We train the network using the Adam optimizer for 25 epochs with a batch size of 128 and a learning rate of 0.00001. To evaluate the FA network, we calculated the FA prediction accuracy as well as the F-score. Our FA network achieves an average accuracy of 91%, and an F-score of 0.89.

C. ENGAGEMENT CLASSIFICATION

To evaluate this stage, an online study was set up in the lab space of the Neuro-Information Technology research group, Otto-von-Guericke University, Magdeburg, Germany. We have conducted on-site experiments with 11 subjects (2 ♀| 9 ♂) aged between 20 and 34 years following the scenario mentioned in Sec. III. Data has been recorded during the experiments using dedicated ROS message data structures and saved in bags, the primary mechanism in ROS for data logging. Since each message comes with a specific timestamp of its generation, all streams can be easily synchronized. An example of the recorded multimodal features is shown in Fig. 7. When the subject's focus of attention deviates from the robot for a while, the model predicts a subject disengagement which is highlighted in the figure by the pink color.

It is hard to annotate a person's engagement with the robot as it is subjective and requires trained annotators. We introduce a benchmark to evaluate our model based on certain engagement and disengagement events that can be easily annotated. During the experiments, the subject was expected to engage with WS2 five times during the whole experiment: one time for registration, three times for interaction, and one time to answer The questionnaires. Also, the subject was expected to engage with WS1 four times for the task execution. Consequently, the subject will have nine engagement and disengagement events.

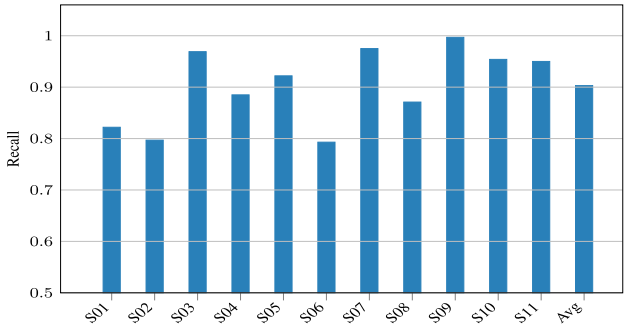


FIGURE 8. Subject-wise and average precision values using our model.

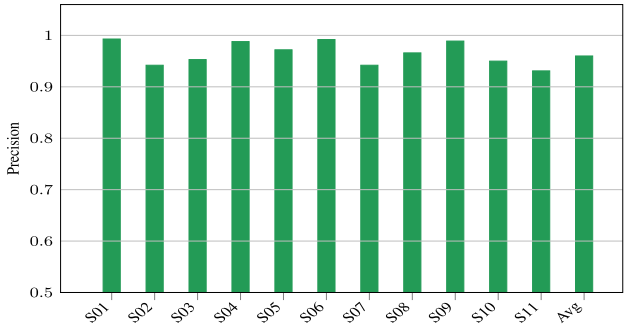


FIGURE 9. Subject-wise and average recall values using our model.

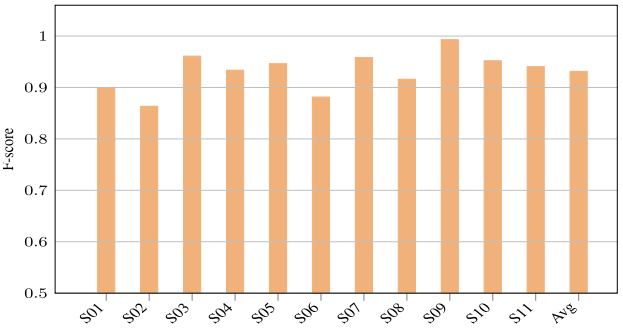


FIGURE 10. Subject-wise and average F-score values using our model.

In addition, we design a way to track the engagement and disengagement actions during the study. If the robot, infers that the user wants to engage, it turns toward the user; otherwise, it does not turn. To assess the performance of our model, data is extracted from the bags and classified into True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN). Based on the classification results, we calculate the model precision, recall, and F-score for each subject separately. Fig. 8, 9, 10 shows the precision, recall, and F-score for each subject separately and the average of all subjects. As can be seen, our model achieves an average of 96%, 90%, and 93% precision, recall, and F-score

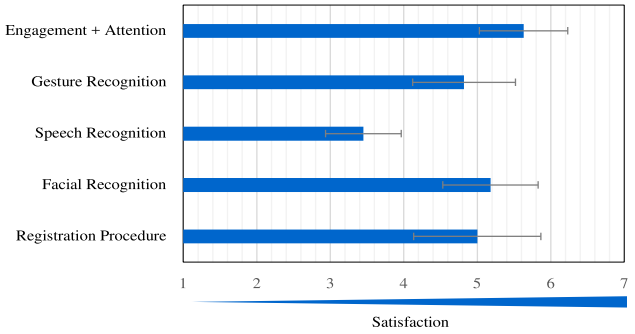


FIGURE 11. Subject rating of individual modules.

respectively. These metrics are calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F = 2 * \frac{precision * recall}{precision + recall} \tag{3}$$

D. SUBJECT EVALUATION

We use a module-specific questionnaire at the end of the experiments to increase the robustness of the evaluation. Subjects were asked to rate all the modules implemented in the study including the engagement process according to their personal satisfaction on a scale of one, very dissatisfied, to seven, very satisfied. The answer of the subjects can give us a better intuition about the performance of the engagement process with the cobot. Fig. 11 shows the average rating of all modules based on the subject's answers. From the figure, the subjects rate the engagement process with the highest value of 5.65 which means that they were more satisfied with the engagement process compared with other modules.

E. PERFORMANCE EVALUATION

The critical factors for the performance evaluation of the overall system are the delays and system response times. Real-time capability is hereby essential. We define real-time as a system responding within a time frame defined by used hardware restrictions, as it is not possible to surpass them. If the system does not add additional delays due to its calculations and algorithms, then it is performing well.

In order to get accurate performance evaluation for our model, we run the experiments again on the recorded videos from the online study and calculate the average execution time of our model and the delay. The Kinect V2 defines the bottleneck for the response times, So it is deployed on a separate PC (Specs: Intel Core i7-8700 @ 3.2GHz, NVIDIA 1080 TI, 32Gb RAM, SSD) in which the video and skeleton feeds are capped at 30 *fps* with average 33.$\bar{3}$) *ms*.

The other three models are deployed on another PC (Specs: Intel Core i7-9800X @ 3.8GHz, NVIDIA RTX 3080, 32Gb RAM, SSD) in which the output features are capped at 47 *fps* with an average of 21.2 *ms*. Table 2 shows the average execution time of individual methods used in the first stage of our model. To summarize, the average execution time for the first stage is equal to 33.3 *ms* as we execute models simultaneously. The other stages are not consuming more than 1 *ms* as it contains simple rule-based policies. Consequently, the total processing time of a frame in our three-stage model is approximately 40 *ms* including 7 *ms* delays.

TABLE 2. Average execution time of individual methods used in the first stage of our model.

Method	Average Time (ms)
Face Detection	5.2
Face Recognition & Tracking	5.5
Gaze	3.5
Head Pose	7
Kinect V2	33

VI. DISCUSSION AND FUTURE WORK

The study conducted has a complex scenario and hardware setup which include a human interacting with two work stations synchronized together using the ROS operating system. It includes a total of 18 engagement events for each subject in an ideal scenario. In addition, extracting features during the experiments is a challenging task due to the lighting conditions, extreme deviation in head pose and gaze angles, and occlusion. However, the performance assessment mentioned above has illustrated the effectiveness of the proposed model in inferring the subject's engagement and disengagement in a multi-person environment. Our model can predict engagement and disengagement in real-time.

Some engagement decisions caused a long delay during the experiments due to the model computational cost and the network overhead. Although two of the subjects were wearing face masks for the whole experiment, our model succeeded to characterize their engagement with reasonable accuracy.

The advantage of our model is that it depends on simple rule-based policies which do not require manual and costly annotations. Furthermore, subjects were more satisfied with the engagement process based on the subject assessment questionnaire.

Due to the limited number of subjects in this study, future work would contain a new study with a large number of subjects to effectively assess the performance of the model. Also, the future study could contain different human-robot interaction scenarios to test the generalization capabilities of the proposed model. Finally, future work should assess the performance of our model in comparison to learning models obtained from the labeled ground truth.

VII. CONCLUSION

In this paper, we present a robust three-stage model for predicting engagement and disengagement in an open-world environment. Our model consists of a combination of deep

learning and rule-based approaches. Firstly, state-of-the-art deep learning models are utilized to extract robust engagement features including gaze, head pose, body posture, and face ID. Secondly, a feedforward neural network is used to estimate the subject's focus of attention using visual data including head pose and gaze direction. In order to handle engagement in a multi-subject interaction, we use the face recognition technique for person identification to track all the subjects in the scene. Further, a rule-based policy is used to match face ID with the body to maintain an efficient interaction between one subject and the cobot. Finally, a rule-based algorithm is adopted in order to make engagement and disengagement decisions (i.e. when and whom to engage with).

To show the robustness of our model, we validate the methods used in each stage of our model. Regarding the final stage, we conduct an online study with subjects doing collaborative tasks with a cobot. Our model achieves an average of 96%, 90%, and 93% precision, recall, and F-score respectively. The experimental results have shown that our model is able to predict real-time engagement state effectively in a multi-subject environment.

REFERENCES

[1] P. A. Lasota, *A Survey of Methods for Safe Human-Robot Interaction*, vol. 104. Delft, The Netherlands: Now, 2017.

[2] A. Hentout, M. Aouache, A. Maoudj, and I. Akli, "Human–robot interaction in industrial collaborative robotics: A literature review of the decade 2008–2017," *Adv. Robot.*, vol. 33, nos. 15–16, pp. 764–799, Aug. 2019.

[3] A. Zacharaki, I. Kostavelis, A. Gasteratos, and I. Dokas, "Safety bounds in human robot interaction: A survey," *Saf. Sci.*, vol. 127, Jul. 2020, Art. no. 104667.

[4] C. Oertel, G. Castellano, M. Chetouani, J. Nasir, M. Obaid, C. Pelachaud, and C. Peters, "Engagement in human-agent interaction: An overview," *Frontiers Robot. AI*, vol. 7, p. 92, Aug. 2020.

[5] L. J. Corrigan, C. Peters, D. Küster, and G. Castellano, "Engagement perception and generation for social robots and virtual agents," in *Toward Robotic Socially Believable Behaving Systems*, vol. 1. Cham, Switzerland: Springer, 2016, pp. 29–51.

[6] K. Kompatsiari, F. Ciardo, D. De Tommaso, and A. Wykowska, "Measuring engagement elicited by eye contact in human-robot interaction," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 6979–6985.

[7] S. Ivaldi, S. Lefort, J. Peters, M. Chetouani, J. Provasi, and E. Zibetti, "Towards engagement models that consider individual factors in HRI: On the relation of extroversion and negative attitude towards robots to gaze and speech during a human–robot assembly task," *Int. J. Social Robot.*, vol. 9, no. 1, pp. 63–86, Jan. 2017.

[8] K. Kompatsiari, F. Ciardo, V. Tikhanoff, G. Metta, and A. Wykowska, "It's in the eyes: The engaging role of eye contact in HRI," *Int. J. Social Robot.*, vol. 13, no. 3, pp. 525–535, Jun. 2021.

[9] I. Leite, M. McCoy, D. Ullman, N. Salomons, and B. Scassellati, "Comparing models of disengagement in individual and group interactions," in *Proc. 10th Annu. ACM/IEEE Int. Conf. Hum.-Robot Interact.*, Mar. 2015, pp. 99–105.

[10] S. M. Anzalone, G. Varni, E. Zibetti, S. Ivaldi, and M. Chetouani, "Automated prediction of extraversion during human-robot interaction," in *Proc. Int. J. Social. Robot.*, 2015, pp. 29–39.

[11] A. Ben-Youssef, C. Clavel, and S. Essid, "Early detection of user engagement breakdown in spontaneous human-humanoid interaction," *IEEE Trans. Affect. Comput.*, vol. 12, no. 3, pp. 776–787, Jul. 2019.

[12] D. Vaufreydaz, W. Johal, and C. Combe, "Starting engagement detection towards a companion robot using multimodal features," *Robot. Auto. Syst.*, vol. 75, pp. 4–16, Jan. 2016.

[13] D. Bohus and E. Horvitz, "Managing human-robot engagement with forecasts and... um... hesitations," in *Proc. 16th Int. Conf. Multimodal Interact.*, Nov. 2014, pp. 2–9.

[14] D. Bohus and E. Horvitz, "Dialog in the open world: Platform and applications," in *Proc. Int. Conf. Multimodal Interfaces (ICMI-MLMI)*, 2009, pp. 31–38.

[15] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. W. McOwan, "Detecting engagement in HRI: An exploration of social and task-based context," in *Proc. Int. Conf. Privacy, Secur., Risk Trust Int. Confernece Social Comput.*, Sep. 2012, pp. 421–428.

[16] A. Ben-Youssef, G. Varni, S. Essid, and C. Clavel, "On-the-fly detection of user engagement decrease in spontaneous human–robot interaction using recurrent and deep neural networks," *Int. J. Social Robot.*, vol. 11, no. 5, pp. 815–828, Dec. 2019.

[17] M. E. Foster, A. Gaschler, and M. Giuliani, "Automatically classifying user engagement for dynamic multi-party human–robot interaction," *Int. J. Social Robot.*, vol. 9, no. 5, pp. 659–674, Nov. 2017.

[18] K. Wegener, W. H. Chen, F. Dietrich, K. Dröder, and S. Kara, "Robot assisted disassembly for the recycling of electric vehicle batteries," in *Proc. CIRP*, vol. 29, Apr. 2015, pp. 716–721.

[19] K. I. Alevizos, C. P. Bechlioulis, and K. J. Kyriakopoulos, "Physical human–robot cooperation based on robust motion intention estimation," *Robotica*, vol. 38, no. 10, pp. 1842–1866, Oct. 2020.

[20] D. Romero, J. Stahre, T. Wuest, O. Noran, P. Bernus, Å. Fast-Berglund, and D. Gorecky, "Towards an operator 4.0 typology: A human-centric perspective on the 4th industrial revolution technologies," in *Proc. Int. Conf. Comput. Indust. Eng. (CIE)*, 2016, pp. 29–31.

[21] I. Maurtua, A. Ibarguren, J. Kildal, L. Susperregi, and B. Sierra, "Human–robot collaboration in industrial applications: Safety, interaction and trust," *Int. J. Adv. Robot. Syst.*, vol. 14, no. 4, 2017, Art. no. 1729881417716010.

[22] C. L. Sidner, C. D. Kidd, C. Lee, and N. Lesh, "Where to look: A study of human-robot engagement," in *Proc. 9th Int. Conf. Intell. User Interface (IUI)*, 2004, pp. 78–84.

[23] K. Li, S. Sun, X. Zhao, J. Wu, and M. Tan, "Inferring user intent to interact with a public service robot using bimodal information analysis," *Adv. Robot.*, vol. 33, nos. 7–8, pp. 369–387, Apr. 2019.

[24] L. Li, Q. Xu, and Y. K. Tan, "Attention-based addressee selection for service and social robots to interact with multiple persons," in *Proc. Workshop at SIGGRAPH Asia (WASA)*, 2012, pp. 131–136.

[25] B. Schuller, R. Müeller, B. Höernler, A. Höethker, H. Konosu, and G. Rigoll, "Audiovisual recognition of spontaneous interest within conversations," in *Proc. 9th Int. Conf. Multimodal Interfaces (ICMI)*, 2007, pp. 30–37.

[26] V. Richter, B. Carlmeyer, F. Lier, S. M. Z. Borgsen, D. Schlangen, F. Kummert, S. Wachsmuth, and B. Wrede, "Are you talking to me?: Improving the robustness of dialogue systems in a multi party HRI scenario by incorporating gaze direction and lip movement of attendees," in *Proc. 4th Int. Conf. Hum. Agent Interact.*, Oct. 2016, pp. 43–50.

[27] D. Strazdas, J. Hintz, A.-M. Felßberg, and A. Al-Hamadi, "Robots and wizards: An investigation into natural human–robot interaction," *IEEE Access*, vol. 8, pp. 207635–207642, 2020.

[28] A. Ben-Youssef, C. Clavel, S. Essid, M. Bilac, M. Chamoux, and A. Lim, "UE-HRI: A new dataset for the study of user engagement in spontaneous human-robot interactions," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, Nov. 2017, pp. 464–472.

[29] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, "Gaze360: Physically unconstrained gaze estimation in the wild," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6912–6921.

[30] V. Albiero, X. Chen, X. Yin, G. Pang, and T. Hassner, "img2pose: Face alignment and detection via 6DoF, face pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7617–7627.

[31] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5203–5212.

[32] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[33] X. Guo, S. Li, J. Yu, J. Zhang, J. Ma, L. Ma, W. Liu, and H. Ling, "PFLD: A practical facial landmark detector," 2019, *arXiv:1902.10859*.

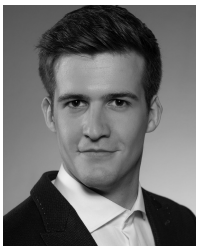
[34] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.

[35] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5525–5533.
[36] S. Chen, Y. Liu, X. Gao, and Z. Han, "MobileFaceNets: Efficient CNNs for accurate real-time face verification on mobile devices," in *Proc. Chin. Conf. Biometric Recognit.* Cham, Switzerland: Springer, 2018, pp. 428–438.
[37] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 87–102.
[38] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Gool, "Random forests for real time 3D face analysis," *Int. J. Comput. Vis.*, vol. 101, no. 3, pp. 437–458, 2013.

AHMED A. ABDELRAHMAN was born in Cairo, Egypt, in 1989. He received the B.Sc. and M.Sc. degrees in electrical and computer engineering from MTC. He is currently pursuing the Ph.D. degree in electrical engineering with the Otto-von-Guericke University Magdeburg. Since 2021, he has been a Research Assistant with the Neuro-Information Technology Research Group, Otto-von-Guericke University Magdeburg. His research interests include computer vision, deep learning, and human–machine interaction.

DOMINYKAS STRAZDAS (Graduate Student Member, IEEE) was born in Vilnius, Lithuania, in 1989. He received the B.Sc. and M.Sc. degrees in mechatronics from Otto-von-Guericke University Magdeburg, where he is currently pursuing the Ph.D. degree in electrical engineering. Since 2017, he has been a Research Assistant with the Neuro-Information Technology Research Group, Otto-von-Guericke University Magdeburg. His research interests include human–machine interaction, especially natural, intuitive, and contact-less communication between humans and robots. He conducted RoSA study, which was awarded with 2020 IEEE Access Best Multimedia Award (Part 2) as Promotional Prize Winner

ALY KHALIFA was born in Cairo, Egypt, in 1984. He received the B.Sc. and M.Sc. degrees in electrical and computer engineering from MTC. He is currently pursuing the Ph.D. degree in electrical engineering with Otto-von-Guericke University Magdeburg. Since 2020, he has been a Research Assistant with the Neuro-Information Technology Research Group, Otto-von-Guericke University Magdeburg. His research interests include computer vision, human–machine interaction, and deep learning—deep face recognition.



JAN HINTZ was born in Brunswick, Lower-Saxony, Germany, in 1996. He received the M.Sc. degree in electrical engineering and information technology from Otto-von-Guericke University Magdeburg, where he is currently pursuing the Ph.D. degree. Since 2018, he has been a Research Assistant with the Neuro-Information Technology Research Group, Otto-von-Guericke University Magdeburg. His research interests include computer vision, image processing, machine learning, and human–machine interaction.

THORSTEN HEMPEL was born in Pinneberg, Schleswig-Holstein, Germany, in 1993. He received the B.S. degree in industrial engineering from the West Coast University of Applied Sciences, Heide, Germany, in 2017, and the M.S. degree in industrial engineering from the Otto-von-Guericke-University of Magdeburg, Germany, in 2019. Since 2019, he has been a Research Assistant with the Neuro-Information Technology Research Group, Otto-von-Guericke University Magdeburg. His research interests include the field of robot vision, robot navigation, and human–robot interaction.

AYOUB AL-HAMADI received the Ph.D. degree in technical computer science, in 2001, the Habilitation degree in artificial intelligence, and the Venia Legendi degree in pattern recognition and image processing from Otto-von-Guericke University Magdeburg, Germany, in 2010. He is currently a Professor and the Head of the Neuro-Information Technology Group, Otto-von-Guericke University Magdeburg. He is the author of more than 380 papers in peer-reviewed international journals, conferences, and books. His research interests include computer vision, pattern recognition, and image processing. See http://www.nit.ovgu.de/ for more details.