https://doi.org/10.31436/iiumej.v24i1.2577

NON-VERBAL HUMAN-ROBOT INTERACTION USING NEURAL NETWORK FOR THE APPLICATION OF SERVICE ROBOT

ZUBAIR ADIL SOOMRO[1], ABU UBAIDAH BIN SHAMSUDIN[1*], RUZAIRI BIN ABDUL RAHIM[1], ANDI ADRIANSYAH[2], MOHD HAZELI[3]

[1] Department of Electronic Engineering,
Faculty of Electrical and Electronic Engineering,
Universiti Tun Hussein Onn Malaysia, Batu Pahat, Malaysia
[2] Universitas Mercu Buana, Jakarta, Indonesia
[3] Move Robotic SDN BHD

**Corresponding author: ubaidah@uthm.edu.my*

(Received: 22nd August 2022; Accepted: 4th December 2022; Published on-line: 4th january 2023)

ABSTRACT: Service robots are prevailing in many industries to assist humans in conducting repetitive tasks, which require a natural interaction called Human Robot Interaction (HRI). In particular, nonverbal HRI plays an important role in social interactions, which highlights the need to accurately detect the subject's attention by evaluating the programmed cues. In this paper, a conceptual attentiveness model algorithm called Attentive Recognition Model (ARM) is proposed to recognize a person's attentiveness, which improves the accuracy of detection and subjective experience during nonverbal HRI using three combined detection models: face tracking, iris tracking and eye blinking. The face tracking model was trained using a Long Short-Term Memory (LSTM) neural network, which is based on deep learning. Meanwhile, the iris tracking and eye blinking use a mathematical model. The eye blinking model uses a random face landmark point to calculate the Eye Aspect Ratio (EAR), which is much more reliable compared to the prior method, which could detect a person blinking at a further distance even if the person was not blinking. The conducted experiments for face and iris tracking were able to detect direction up to 2 meters. Meanwhile, the tested eye blinking model gave an accuracy of 83.33% at up to 2 meters. The overall attentive accuracy of ARM was up to 85.7%. The experiments showed that the service robot was able to understand the programmed cues and hence perform certain tasks, such as approaching the interested person.

ABSTRAK: Robot perkhidmatan lazim dalam banyak industri untuk membantu manusia menjalankan tugas berulang, yang memerlukan interaksi semula jadi yang dipanggil Interaksi Robot Manusia (HRI). Khususnya, HRI bukan lisan memainkan peranan penting dalam interaksi sosial, yang menonjolkan keperluan untuk mengesan perhatian subjek dengan tepat dengan menilai isyarat yang diprogramkan. Dalam makalah ini, algoritma model perhatian konseptual yang dipanggil Model Pengecaman Perhatian (ARM) dicadangkan untuk mengenali perhatian seseorang, yang meningkatkan ketepatan pengesanan dan pengalaman subjektif semasa HRI bukan lisan menggunakan tiga model pengesanan gabungan: pengesanan muka, pengesanan iris dan mata berkedip. . Model penjejakan muka telah dilatih menggunakan rangkaian saraf Memori Jangka Pendek Panjang (LSTM), yang berdasarkan pembelajaran mendalam. Manakala, pengesanan iris dan mata berkelip menggunakan model matematik. Model mata berkelip menggunakan titik mercu tanda muka rawak untuk mengira Nisbah Aspek Mata (EAR), yang jauh lebih dipercayai berbanding kaedah sebelumnya, yang boleh mengesan seseorang berkelip pada jarak yang lebih jauh walaupun orang itu tidak berkelip. Eksperimen yang dijalankan untuk pengesanan muka dan iris dapat mengesan arah sehingga 2 meter. Sementara itu, model berkelip mata yang diuji memberikan ketepatan

https://doi.org/10.31436/iiumej.v24i1.2577

83.33% sehingga 2 meter. Ketepatan perhatian keseluruhan ARM adalah sehingga 85.7%. Eksperimen menunjukkan bahawa robot perkhidmatan dapat memahami isyarat yang diprogramkan dan seterusnya melaksanakan tugas tertentu, seperti mendekati orang yang berminat.

KEYWORDS: *Attentive Recognition Model (ARM), Long-Short-Term Memory (LSTM), Human Robot Interaction (HRI), Eye Aspect Ratio (EAR)*

1 INTRODUCTION

As service robots are starting to coexist with humans, they ought to deal with two fundamental aspects: sturdy navigation in cluttered environments [1, 2] and effective human-robotic interaction (HRI). There have been many research conducted in navigation area whereas, a significant amount of problems in HRI field need to be addressed. Therefore, there are two primary forms of communication in robotics for establishing an accurate interaction between a robot and a human being within physical and psychological contexts. Human verbal interaction is expressed using thoughts, ideas, and feelings through spoken or written language. Meanwhile, in kinesics, nonverbal communication is defined through body movements, positioning, facial expressions, and gestures [17]. In particular, this research is based on head movement and oculesics, a subcategory of kinesics, which is the study of eye movement, behaviour, gaze, and eye-related nonverbal communication. An important perceptual variable in the analysis of nonverbal cues is the focus of attention, which indicates the person is attending to something and is often used to convey interpersonal intent. It often happens that people turn toward their focus of attention, thereby physically expressing their attention by means of head orientation. In most cases, head orientation is directly related to the visual gaze estimate since the perceived gaze direction is determined by the orientation of the head. Inspired by this human behaviour, an Attentive Recognition Model (ARM) is proposed that corresponds to social cues, namely, face tracking, iris tracking and eye blinking, characterised by effective non-verbal human interaction.

In the prior method for attentiveness architecture, body components were represented as cylindrical shapes using the Top View Re-projection (TVR) concept [3]. The hypothesis with the highest score provides the presumed pose and the location of the joints after the pose has been measured against a scoring system. Meanwhile, another method uses mutual gaze [4] as a social cue by making iCub capable of recognizing eye contact events while interacting online with a human partner. Other than that, Hand Gesture Recognition in [5] was also considered one of the non-verbal cues used to get the attention of a robot. The methods in the literature are limited to the use of stationary robots and are not considered to be a combined model of different cues. Therefore, this paper will focus on the face tracking, iris tracking and eye blinking cues, which are mainly important because they develop social awareness. Furthermore, both portrayals of robots and embodied robots activate social cognitive mechanisms that rely on interpreting others' gaze direction, such as gaze cueing. The significance of a glance in a particular direction can only be completely understood if the gaze's objective and the agent's mental state are known in relation to that target [6]. Author in [7] presented a study on social cues being taking into account to enable the robot by deciding socially, at which human it should direct its gaze.

The author in [8] used a performance improved eye tracking system for a more efficient human-robot collaboration than a comparable head tracking approach. Paper [9] introduced an approach for accurate and robust eye center localization by using image gradients. This method yields low computational complexity and is invariant to rotation and linear changes in

https://doi.org/10.31436/iiumej.v24i1.2577

illumination. It was also tested in [10], where a Partially-Observable Markov Decision Process (POMDP) was proposed to plan the navigation in the dynamic (crowded) environment by using sensor fusion and filtering techniques.

In addition, the previous method in [11] for the eye blinking model used vertical and horizontal lines to calculate blinking ratio, which limited its capability to detect blinks at a further distance. Moreover, it was also added to help the robot show more realistic eye gazes. Nevertheless, the overall accuracy of the system is not precise enough for different light intensity environments for altogether algorithms.

On the other hand, the previous method in [12, 13] for the face tracking model uses Support Vector Machine (SVM) and Constrained Local Models (CLM) to train the data, respectively. Although the former option provides enough precision in gaze detection, the latter avoids any tedious calibration procedure and makes it possible to interact with people with no prior preparation. Thus, a new Long Short-Term Memory (LSTM) neural network was considered to train the 468 extracted keypoints from the face. This network is an extension of RNN, designed to avoid the long-term dependency issue. It can also remember data for long periods. This happens as it allows back-propagation through time by connecting events that appear far apart in the input data without their weight being diluted between the forget gate, input gate and output layers.

The main objectives of this work are as mentioned below:

- To propose an attentive gaze algorithm by fusing iris, face tracking and eye blinking data.
- To develop face tracking model using LSTM neural network.
- To optimize previous eye blinking model using a new variable
- To validate the feasibility of the method in real world for service robot.

2 PROPOSED ARM METHOD

The service robots in a café operate in close proximity with humans, which raises social behavioural skills typical of human–human interactions, which is still a challenge [4]. This research proposes an attentive gaze algorithm by estimating the face direction, iris direction and eye blinking as a fundamental social cue in face-to-face interactions. These models were integrated into the service robot to maximize its capability to recognize the stare of a person, which potentially gives it an advantage in communicating by reducing false negative errors. The data of trained LSTM keypoints in 4.4 were used to detect the possibility of the face posing classification. Based on the face detected in a person, the iris and eye blinking models were implemented. The ARM algorithm was tested in 4.5 by facing straight at the camera. When the face was facing away from the camera, the algorithm was not able to detect the attentiveness from the person, hence there was no response from the system. When a person showed some cues, such as facing straight at the camera with direct eye contact by aligning the iris position and no eye blinking, the system gave a response of '!!ATTENTIVE!!' as shown in Fig 1. The experimentation for face, iris tracking and eye blinking was conducted in Sections 4.2, 4.3 and 4.4, respectively. This shows that the attentive gaze algorithm can be used to gain a person's interest. This combined model was important as it could execute all of the important cues altogether without having multiple files.

https://doi.org/10.31436/iiumej.v24i1.2577

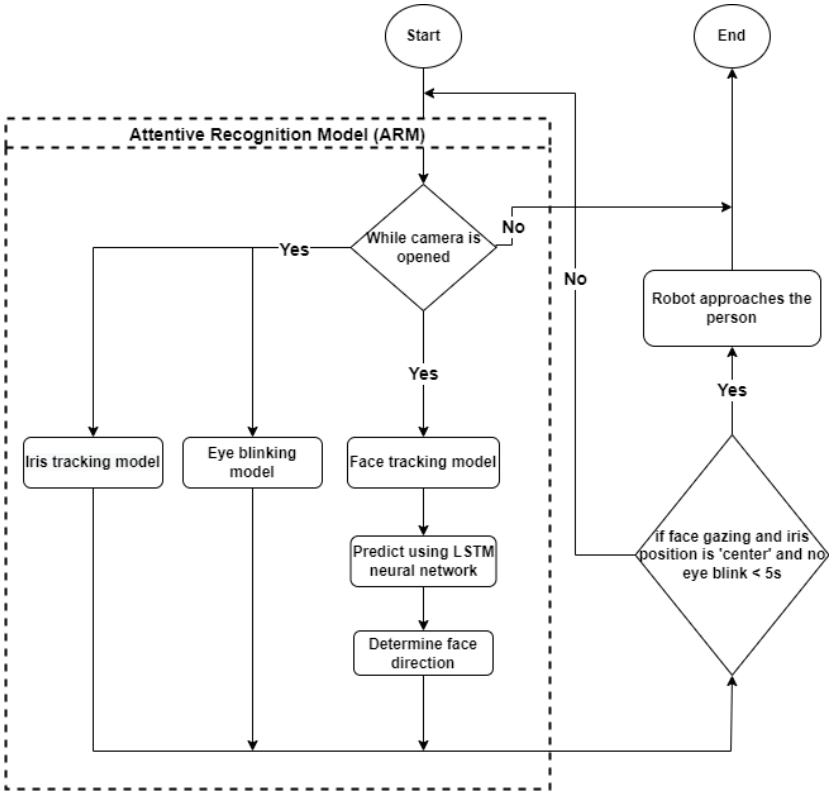

Fig. 1: Overall flowchart of ARM

3 METHODOLOGY

3.1 Data collection

A participant was recruited for data collection (age = 23) for a face tracking model using a Lenovo S410p laptop camera. A total of 20 number of frames were extracted from the 20 raw video sequences for each of the action namely right, left and center. The data collection was conducted at the Universiti Tun Hussein Onn Malaysia.

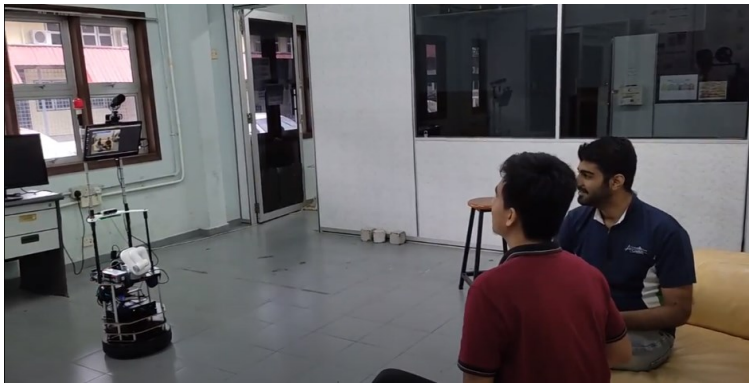

Fig. 2. Human interaction with robot

3.2 Environment setup

Hussein, a service robot developed by the UTHM RoboCup team, was used in this experiment as shown in Fig. 2. It was located in the middle of the room and was surrounded by chairs and tables to look like a café environment. A Logitech C920 high-definition (HD) webcam was used for tracking, which provides a clear image up to 3 meters away. It was

https://doi.org/10.31436/iiumej.v24i1.2577

mounted on the robot at a height of 130 cm from the ground to detect different heights of a person's face. The developed software modules were integrated within the ROS environment, which was needed to initiate the robot's navigation system.

3.3 Task

A total of 7 participants (mean age = 22.5) were chosen to test the ARM algorithm. They were asked to sit in front of the robot at a distance of around 2 meters and establish mutual gaze.

3.4 Measurement

The measurements of this experiment have been separated into two categories:

1. Testing of each models in indoor and outdoor environment
2. Accuracy of ARM detection

3.5 Face tracking

For the active eye contact recognition, the face location is tracked. For this purpose, state-of-the-art real-time face mesh approach was utilized from Google Research, which is implemented in MediaPipe-framework. This approach differs from [15] by using machine learning to infer 3D surface geometry that estimates 468 3D face landmarks in real-time from a single frame and facilitates fast and robust requiring only a single camera input and no separate depth sensor as shown in Fig 3. It is able to utilize lightweight model designs and CPU acceleration across the pipeline to achieve real-time performance, which is crucial for live experiences [16]. This ML pipeline operates on the captured video using OpenCv library and computes face positions and a 3D face landmark model that uses those locations to predict the approximate surface geometry via regression. This also have advantage of greatly reducing the requirement for typical data augmentations such as affine transformations, which include rotations, translations, and scale modifications.

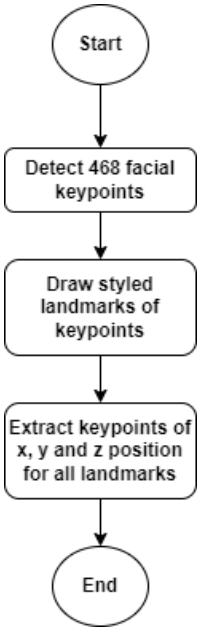

Figure 3: Flowchart for face tracking model

Theoretically, the line linking the eye centres with the rectangle's horizontal axis is aligned by rotating a facial rectangle using the bounding rectangles produced by the camera's frame. To create the input for the mesh prediction neural network, this is then clipped and resized. A

https://doi.org/10.31436/iiumej.v24i1.2577

vector of 3D landmark coordinates can then be generated by the model and mapped back into the original image's coordinate system. The facial rectangles of a real-world frame were covered using a 3D Morphable Model (3DMM) technique, and the ground truth vertex coordinates are now readily available thanks to a previously established relationship between the 468 mesh points and a subset of 3DMM vertices.

This mesh model pays special attention to semantically significant face regions, resulting in more accurate landmark predictions around the mouth, eyes, and irises at the cost of average computation as it resulted in 4.2. The start of algorithm for the face tracking process was detected for specified class and gathering landmarks keypoints. Those keypoints were saved in their respective classes such as ‘center’, ‘left’ and ‘right’.

3.6 Iris tracking

Meanwhile it was known that the horizontal iris diameter of the human eye is fairly constant throughout the population at 11.7±0.5 mm [17]. Due to restricted processing resources, fluctuating lighting conditions, and the presence of occlusions, such as people squinting, it was a difficult assignment to complete. Moreover iris tracking had 10 additional iris landmark which gave accurate estimation for features affecting the iris, pupil, and eye contours in real time using just a single RGB camera and no specialist hardware [18]. It was also able to utilize for determining the metric distance of the camera to the user with relative error less than 10%. Fig. 4 displays a face mesh containing over 468 main keypoints which is represented by the dots (junction between lines).

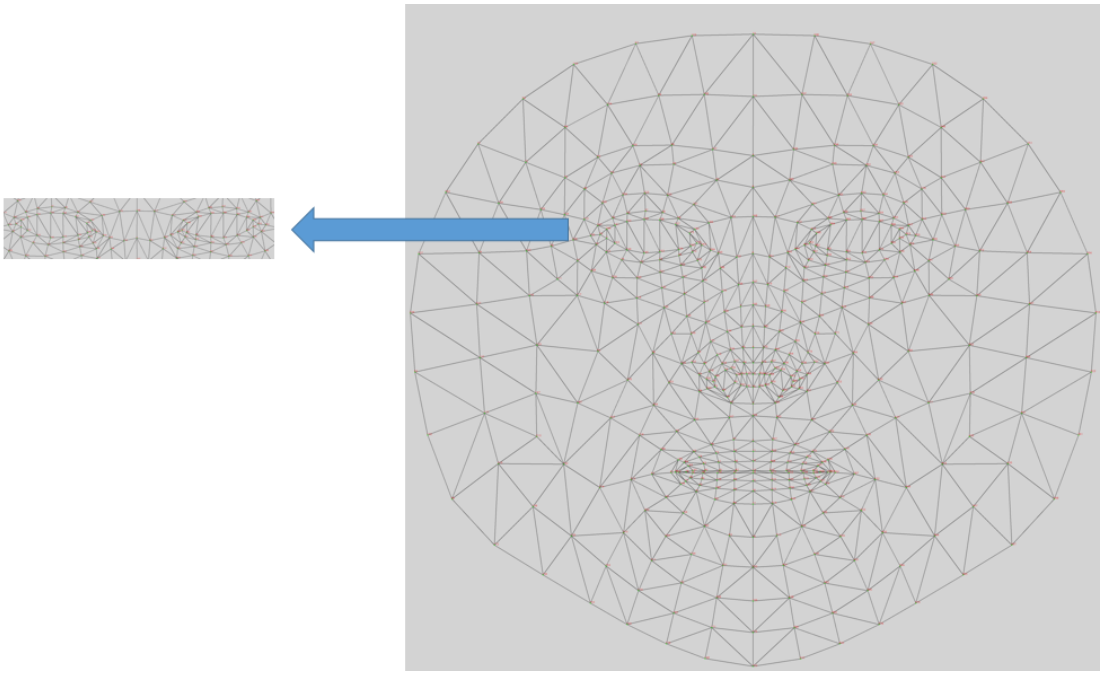

Fig. 4. Eyes landmarks with marked numbers

This was achieved by using the Euclidean distance formula as shown in equation 2.1 where each point reflects the mesh points with notations and corresponding descriptions in Table 1:

$$euclidean\ distance = \sqrt[2]{(x2 - x1)^2 + (y2 - y1)^2} \tag{2.1}$$

https://doi.org/10.31436/iiumej.v24i1.2577

Table 1: Notations and corresponding descriptions

Symbol	Description
x1	x-coordinate of point1
y1	y-coordinate of point1
x2	x-coordinate of point2
y2	y-coordinate of point2

The iris model predicts both eye landmarks (along the eyelid) and iris landmarks (along the iris contour) from an image patch of the eye region. An array of mesh points such as 33, 133, 362 and 263 were generated for x and y coordinates of eyes which was based on the frame size of image captured by the camera as shown in Fig 5. These mesh points then provided the position for left and right eyes for x and y position coordinates respectively to extract the radius and center as resulted in 4.3.

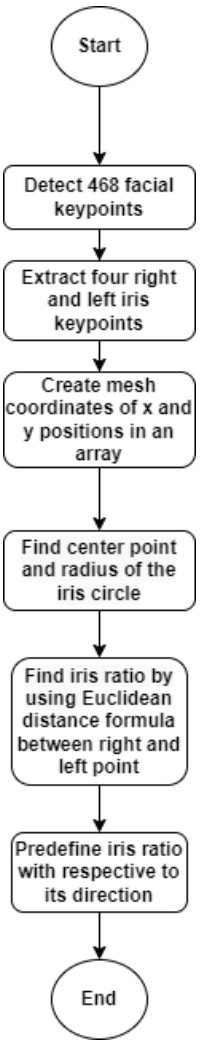

Fig. 5: Flowchart for iris tracking model

3.7 Eye blinking

Although blinking may appear to be an unconscious habit, new research by Paul Hömke and colleagues reveals that when humans are conversing, they unconsciously receive eye blinks as nonverbal indicators [14]. Furthermore, research have revealed that blinks frequently occur during natural pauses in discourse. Hömke wondered if, like nodding one's head, a movement

https://doi.org/10.31436/iiumej.v24i1.2577

as little and subconscious as blinking may work as verbal feedback. Speakers were able to detect a minor difference between short and long blinks, with longer blinks prompting significantly shorter responses from volunteers. More broadly, the discovery may help us understand the origins of how humans communicate their mental states, which has evolved into an important component of everyday social interactions. In this algorithm, a different approach have been used to calculate for eye blinking.

The right and left eyes coordinates were used to find vertical ratio of right and left eye respectively as shown in Fig. 6. As to differentiate this method with previous ways, it was by using the random coordinate landmarks which was in the nose area to get overall ratio. The vertical distance between those coordinates was calculated using the Euclidean distance formula. The calculated blink ratio was utilized by comparing it with a threshold of 0.7 which was determined by using trial and error method as experimented in 4.4.

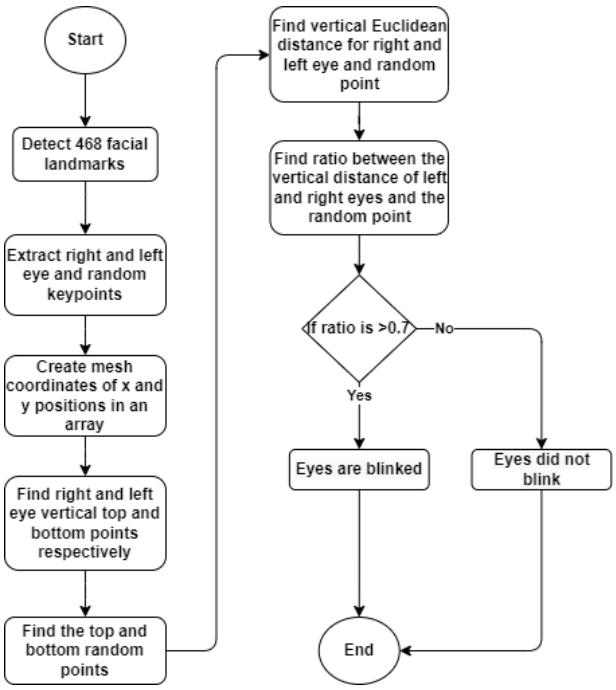

Fig. 6: Flowchart for eye blinking model

3.8 Long-Short-Term Memory (LSTM) Neural Network

In a crowded environment, datasets for people cues detection to identify a large number of human motion data [18] increases linearly and often difficult to interpret due to involvement of sequential data. LSTM is a method capable of learning order dependence in sequence prediction problems by enabling persistent error back-propagation within its inner memory cells. It remembers its input, due to an internal memory, which makes it perfectly suited for machine learning problems that involve sequential data. This will help to reduce the computational costs by reducing number of dimensions of the feature space by extracting a subspace that describes the data best [19]. Hence, reducing number of training data. This study uses LSTM neural network model to process and identify a face direction position. For example, the model proposed in the study is divided into three directions, namely, right, left and center. These directions are used to distinguish the attentiveness of the system function. The collected data was diversified into right, left and center. To train the model, the dataset used in this experiment was based on the personal dataset collection, which is conducive to reflecting the accuracy of LSTM neural network model. The tactile model was trained using sequences of images. It employs three LSTMs to shape temporal data by progressive

https://doi.org/10.31436/iiumej.v24i1.2577

codification of its vector. The output classification probability distribution was provided by a fully-connected layer with 3 neurons and a Softmax function after the second LTSM. A modified LSTM neural network was constructed which consist of LSTM, dropout and dense layers as shown in Fig. 7.

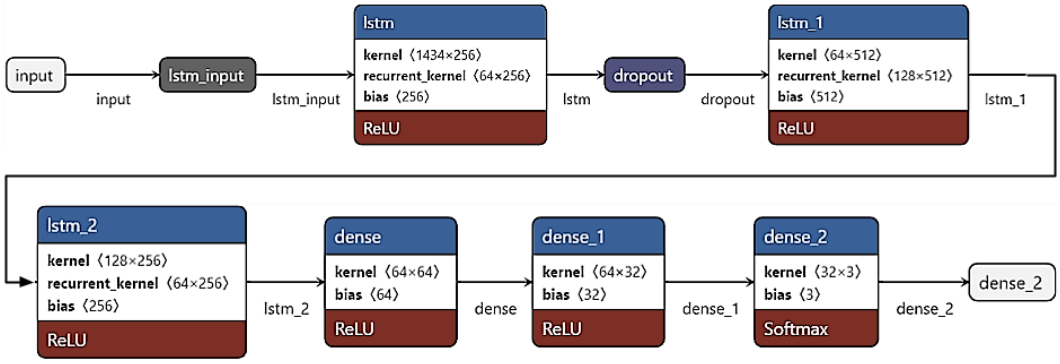

Fig. 7. LSTM neural network architecture

The data was spliced into training and testing keypoints which made it easier to evaluate the trained data. For this case, the original data for each category have been split into 30% of testing and 70% of training. The LSTM neural network trained the data for all of the classes according to the right, left and center face movement classification in 4.5. This type of neural network was powerful as it was capable to create a more descriptive and abstract result to let the network learn all sorts of features from scratch, by arranging them in layers, the network can recycle/reuse features such that low-level features combine to form mid-level features and mid-level features combine to form high-level features.

4 RESULTS AND ANALYSIS

4.1 Face tracking experiment and result

The face detection model was developed to execute the face tracking recognition as cues for attentive model for specified face mesh annotations. This was able to produce more nodes on face which can easily be observed by the joints in tessellation, while previous method [16] consists of smaller number of landmarks which were used for cues to be trained which in this case makes it to have less precision. This was tested in both indoor and outdoor conditions for 2 meters as you can see in the Fig. 8 below.

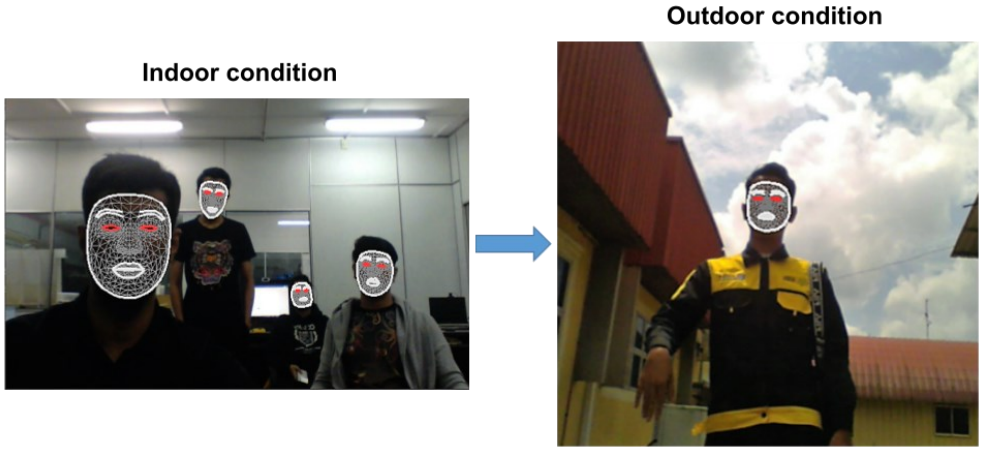

Fig. 8. Face detection model

https://doi.org/10.31436/iiumej.v24i1.2577

From the experiment, it was shown that the model was able to detect the multiple face mesh with limitation roughly up to 2 meters in both indoor and outdoor environment. This could be due to the fact that it was affected by the luminance in the room preventing it to recognize the people. Even though, for outdoor condition the light intensity was higher than the indoor, the model was detected due to the camera's capability to auto calibrate lens adjustment which allows it to adjust the light entering its aperture and detect a person's face. The face tracking model was limited by the amount of data needed to be trained, therefore it is suggested to use larger dataset to have better prediction. This is summarized in the following Table 2.

Table 2: Overall face detection

Participant	Environment	Estimated distance of detection (m)
1	Indoor/Outdoor	2
2 (wearing spectacles)	Indoor/Outdoor	2
3 (wearing spectacles)	Indoor/Outdoor	2

4.2 Iris tracking experiment and result

As there are people in real life who face straight forward but at the same time look the other way which causes the unnaturalness in communication with the robot therefore, an iris tracking model is introduced to create a natural eye to eye contact feature. This algorithm was proposed to extract the iris landmark from the specified face mesh annotations. The ratio for determining the iris positions are classified as in following Table 3.

Table 3: Iris position based on specified ratio

Ratio	Iris position
0 – 0.42	Right
0.42 – 0.57	Center
0.57 – 1.0	Left

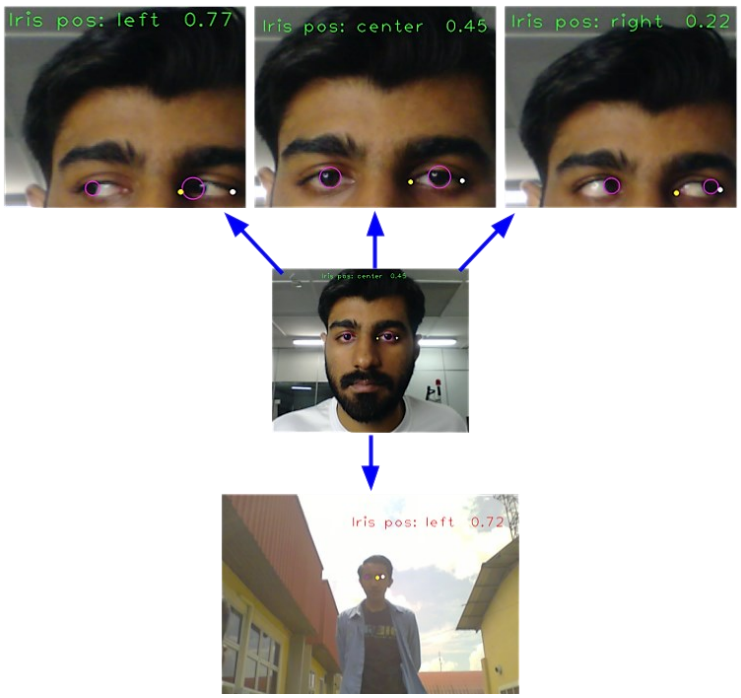

Fig. 9. Iris tracking model

https://doi.org/10.31436/iiumej.v24i1.2577

These values were determined by conducting several trials in both indoor and outdoor environments as shown in Fig. 9. This was calculated using the Euclidean distance equation. The model was able to detect the participant's iris at 2 meters from the camera.

4.3 Eye blinking experiment and result

The eye blinking was tested for squinting and blinking action to test the model's robustness as shown in Fig 10. The state of eye being closed shows that the person was blinking.

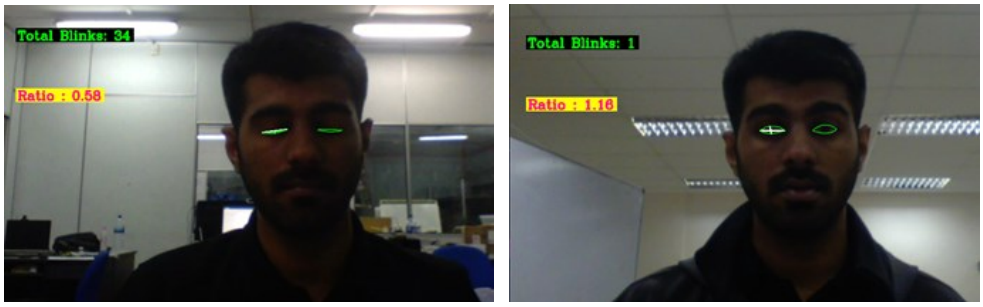

Fig. 10. Eye blinking model

The data also consisted of different positions of face during squinting so that the camera was able to take record of those eyes movement keypoints at most of the positions in camera's stereo vision. The vertical distance used for right, left eyes and random landmarks at nose area is illustrated in Fig 11. These points were chosen due to their availability in such position that can give maximum distance value to get the correct distance measurement. Thus enabling by giving flexibility to test the model for further distance.

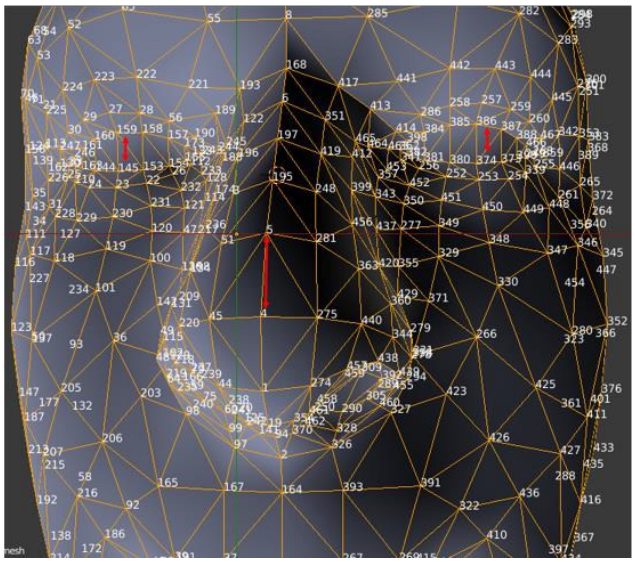

Fig. 11. Vertical distance measurement [14]

From the Fig. 12, it shows that even when a person was standing at farther distance from the camera the ratio is roughly same as of sitting in front of camera. This outlines that the vertical distance between the 2 landmarks of eyes have similar observation in terms of value compared to the vertical distance between the 2 random landmarks chosen. This mathematical method enabled the algorithm to differentiate between eyes open and closed even at further distance.

https://doi.org/10.31436/iiumej.v24i1.2577

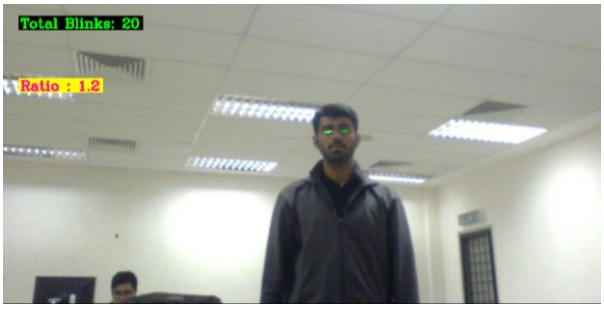

Fig. 12. Eyes open at 2 meters distance

Additionally, participants 2 and 3 wore spectacles to test the capability of the eye blinking model with respect to lenses distortion detection as shown in Fig 13. The previous method [14] of ratio calculation had some drawbacks which was noticed when a person moved further away from the camera, the ratio decreases causing the blinking counter to continuously moves. This was because when a person moves away from camera, the pixels that covers the eyes reduces which causes the distance to be reduced making the counter to increment.

Fig. 13: Testing participant 2 and 3

Table 4 explains the statistics on the prediction and true set for each participants. Statistics for participant 1 on the prediction set show that the number of eyes continue to open for 2 samples, and the number of blinks was 2 samples for EAR threshold 0.7. However, statistics on the test set describe the number of eyes continue to open was 2 samples, and the number of blinks was 2 samples. Furthermore, participant 2 and 3 had eyes continue to open for 2 samples, and the number of blinks was 2 samples for EAR threshold 0.7. Meanwhile, during prediction of blinking only 1 sample was present, and predicted eyes open was 2 samples.

Table 4: Statistics on the prediction and true set for each participant

Participants	True eyes open	Predicted Eyes remains open	True eyes blink	Predicted eyes blink	EAR
1	2	2	2	2	0.7
2	2	2	2	1	0.7
3	2	2	2	1	0.7

Based on the table, a confusion matrix was drawn as shown in Table 5. This experiment exhibited an accuracy of 83.33% by using calculation as shown in equation 3.1 which justifies its practicality. Based on this, it had been concluded that, the experiment achieves the best performance for all datasets.

https://doi.org/10.31436/iiumej.v24i1.2577

Table 5: Confusion matrix for 3 participants

	True positive	True negative
Predicted positive	6	2
Predicted negative	0	4

$$ACC = \frac{(TP + TN)}{(TP + TN + FP + FN)} \times 100\% = \frac{(6 + 4)}{(6 + 4 + 2 + 0)} = 83.33\% \quad (3.1)$$

The F1 score of this method was calculated as shown in equation 3.2 which resulted in 0.86. For easier comparison, this was later compared with the previous datasets by extracting their precision-recall value and tabulating them as shown in Table 6.

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (3.2)$$

$$= \frac{6}{6 + \frac{1}{2}(2 + 0)} = 0.86$$

Table 6: Results obtained for blink detection using the proposed method

Dataset	ZJU	TF (Talking Face)	Our
Precision	0.98	1.0	0.75
Recall	0.898	1.0	1.0
F1	0.94	1	0.86

The accuracy of the blinking model for the proposed method were compared with previous methods as shown in Table 7.

Table 7: Results obtained for blink detection using the proposed method

Methods	Accuracy (%)
Proposed method	83.33
Dlib+CNN	97

Based on the results, it was concluded that the percentage for accuracy dropped due to the person was wearing glasses or the effect from outdoor environment. As the recall was able to get full percentage, it showed that the eye blinking model had the ability to correctly predict the positives out of actual positives. The glasses might have caused the ratio calculation to be off by a certain margin which leads to misreading in eye blinking cue. Other than that, the response time for the camera to be able to detect a person's face took a longer period which resulted in less precise. This outlines, even eyes are capable to be detected in different light conditions due to the capability of the camera's auto calibrating lens adjustment which allows it to detect a person's face and eyes.

https://doi.org/10.31436/iiumej.v24i1.2577

4.4 Long-Short-Term Memory (LSTM) Neural Network experiment and result

The face tracking model dataset was trained with 97% of accuracy and 3% of loss as shown in Figure 14 and 15. It was seen that the error was propagated at 30 to 40 steps throughout the entire network to compute gradients with respect to inputs early in a long sequence. Such models will often overfit on the training set and lose generalizability and accuracy on the test set. Therefore, a dropout regularization with a dropout probability of $p = 0.5$ had been employed in the output layer and between recurrent layers to combat this. Dropout regularization independently sets each weight in consideration to zero with probability p. In response to this, the network cannot rely on a few weights per-example to predict an outcome, lest those weights get pruned in a training step. The model is thus forced to employ many weights to process and predict each example, reducing overfitting. It was deliberately avoided using of dropout on weights between time steps, as doing so effectively eliminates long-range memory

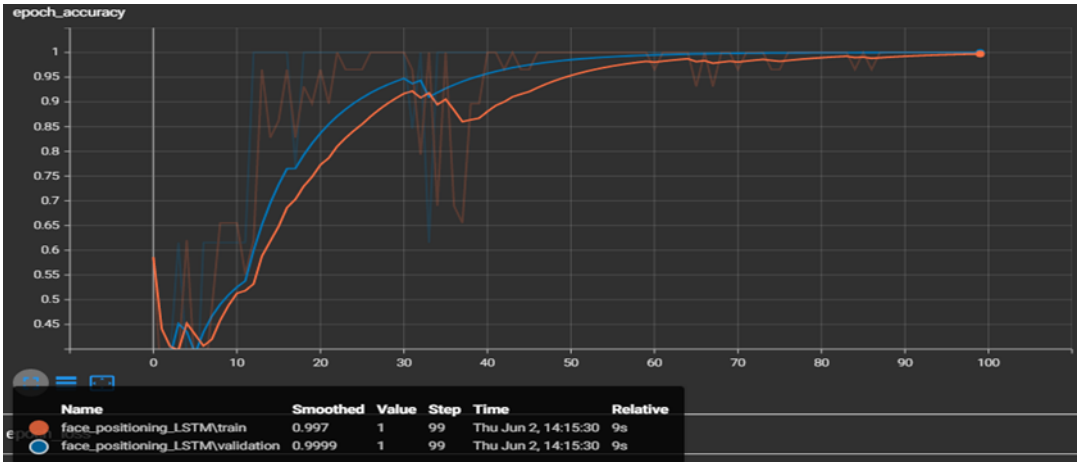

Fig. 14. Training and validation accuracy graph

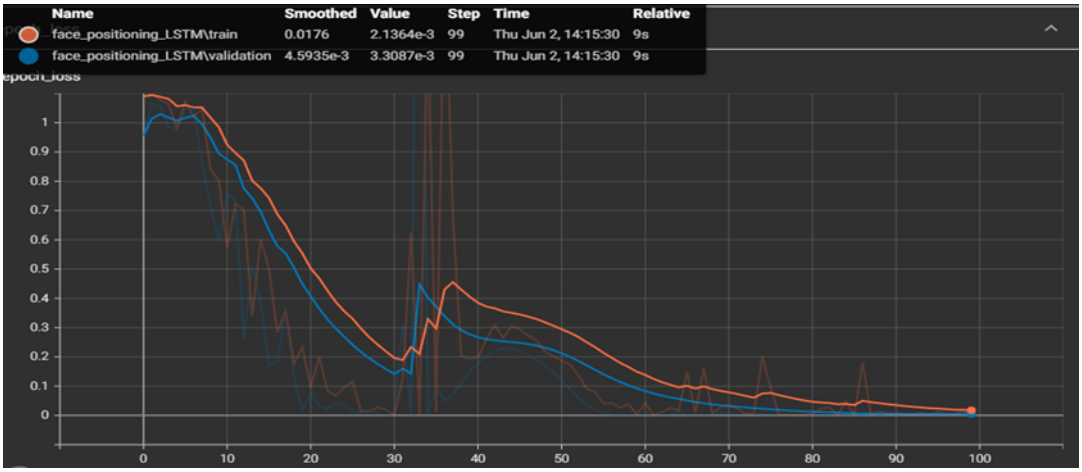

Fig. 15. Training and validation loss graph

4.5 ARM experiment and result

The overall coding for ARM was able to be constructed by combining the mentioned subsection 3.4, 3.5 and 3.6 together. Table 8 demonstrates that the attention mesh runs 80.6% faster than the cascade of separate face and region models on a typical modern mobile device. This performance had been measured using the inference time. Based on this, it could be due

https://doi.org/10.31436/iiumej.v24i1.2577

to during separate execution of the program, it had to run the same coding in other python file rather than using the same line of code form previous saved history such as for the cascaded version of algorithm.

Table 8: Performance on models

Model	Inference Time (ms)
Face mesh tracking	113.98
Iris tracking	37.9
Eye blinking	3.997
Total	155.877
Cascade (ARM)	125.65

7 participants were used to test the system response as tabulated in Table 9. The confusion matrix was formed which showed that the system had 85.7% of accuracy which was due to only one of the participant which had been predicted wrong as shown in equation 3.3. This could be due to the fact that the position of the person was in the right side of the video frame. Therefore, the face was not detected as aligned with the camera's stereo vision.

Table 9: The confusion matrix

	Predicted positive	Predicted negative
Actual positive	4	1
Actual negative	0	2

$$ACC = \frac{(4+2)}{(4+2+0+1)} \times 100\% = 85.7\% \tag{3.3}$$

Recall (R) indicates the ratio of the positive samples which were correctly classified, and presented in equation 3.4.

$$R = \frac{4}{(4+1)} = 0.8 \tag{3.4}$$

Precision (P) as in equation 3.5 was the proportion of actual positive instance in the samples which were classified as positive samples.

$$P = \frac{4}{(4+0)} = 1 \tag{3.5}$$

Intersection over Union (IoU) was calculated for a standard performance measure for the face segmentation problem. IoU represented the overlap between the candidate face mesh and the ground truth face mesh, namely, the ratio of their intersection and union. This showed that, the value was greater due to the closer correlation. Given a set of images, the IoU measurement gave the similarity between the predicted area of the candidate's face mesh presented in the set of images and the ground truth area, which was defined by equation 3.6.

https://doi.org/10.31436/iiumej.v24i1.2577

$$IoU = \frac{area\ of\ overlap}{area\ of\ union} = \frac{4}{(4+0+1)} = 0.8 \quad (3.6)$$

Mean average precision (mAP) for a set of classes was the mean of average precision as in equation 3.7, where N denoted as the number of classes and C meant the class.

$$mAP = \frac{\sum average\ precision}{N(classes)} = \frac{1}{3} = 0.33 \quad (3.7)$$

The accuracy of the ARM was compared with previous methods as shown in Table 10.

Table 10: Results obtained for ARM

Methods	Accuracy (%)
Proposed method	85.7
Scalable HMM	76
OpenPose [5]	97

Based on the result, the accuracy of proposed method was positioned at the middle compared to Scalable HMM and OpenPose. The OpenPose method scored the highest which could be due to its limitation of using eye blinking model to gain attention for attentive model. The literature experiment was conducted based on the study of eye contact. Other than that, it could also be the effect from light which caused the accuracy to drop. This method lacks the iris tracking and eye blinking detection of multiple people at the same time in one frame which is essential for the robot in a crowded place.

5 CONCLUSIONS

In this paper, face tracking, iris tracking and eye blinking models were combined to propose an attentive model known as ARM to improve a robot's attention model performance in determining the most attentive person and prioritizing people based on attentiveness. Moreover, the accuracy of attentiveness prediction was evaluated at different light intensities in order to validate the feasibility of these methods in the real world. This project was expected to provide the best HRI experience with low computational complexity and to be invariant to rotation and linear changes in illumination. Experimentation results concluded that face tracking and an eye blinking model were achieved successfully. This was validated by applying the program inside the service robot and testing it in the real world. Hence, it could be operated in a café environment on a trial basis.

In the future, the current algorithm will be modified to enable it to detect all the important cues of various people at the same time. Additionally, some modifications to the face positioning training data would be done to increase its detection ratio for people at a further distance from the camera.

Acknowledgement: Communication of this research is made possible through monetary assistance by Universiti Tun Hussein Onn Malaysia and the UTHM Publisher's Office via Publication Fund E15216

REFERENCES

[1] Haider, MH, Wang Z, Khan AA, Ali H, Zheng H, Usman S, Kumar R, Bhutta MUM, Zhi P. (2022) Robust mobile robot navigation in cluttered environments based on hybrid adaptive neuro-fuzzy inference and sensor fusion. Journal of King Saud University-Computer and Information Sciences. https://doi.org/10.1016/j.jksuci.2022.08.031

[2] Hacene N, Mendil B. (2021) Behavior-based autonomous navigation and formation control of mobile robots in unknown cluttered dynamic environments with dynamic target tracking. International Journal of Automation and Computing, 18: 766-786. https://doi.org/10.1007/s11633-020-1264-x

[3] Sigalas M, Pateraki M, Trahanias P. (2015, July) Visual estimation of attentive cues in HRI: the case of torso and head pose. International Conference on Computer Vision Systems, vol (9163): pp 375-388. https://doi.org/10.1007/978-3-319-20904-3_34

[4] D'Eusanio A, Simoni A, Pini S, Borghi G, Vezzani R, Cucchiara R. (2020, November) A Transformer-Based Network for Dynamic Hand Gesture Recognition. 2020 International Conference on 3D Vision (3DV): pp. 623-632. https://doi.org/10.1109/3DV50981.2020.00072

[5] Lombardi M, Maiettini E, De Tommaso D, Wykowska A, Natale L. (2022) Toward an Attentive Robotic Architecture: Learning-Based Mutual Gaze Estimation in Human–Robot Interaction. Frontiers in Robotics and AI, 9, 770165. https://doi.org/10.3389/frobt.2022.770165

[6] Hömke P, Holler J, Levinson SC. (2018) Eye blinks are perceived as communicative signals in human face-to-face interaction. PLoS ONE, 13(12): e0208030. https://doi.org/10.1371/journal.pone.0208030

[7] Keiling H. (2019, February 28) 9 Types of Nonverbal Communication and How To Understand Them. Indeed Career Guide [https://www.indeed.com/career-advice/career-development/nonverbal-communication-skills#:~:text=Nonverbal%20communication%20is%20important%20because]

[8] Khan ZH, Siddique A, Lee CW. (2020) Robotics Utilization for Healthcare Digitization in Global COVID-19 Management. International Journal of Environmental Research and Public Health, 17(11): 1-23. https://doi.org/10.3390/ijerph17113819

[9] Timm F, Barth E. (2011) Accurate eye centre localisation by means of gradients. VISAPP 2011 - Proceedings of the International Conference on Computer Vision Theory and Application, 125-130.

[10] Li J, Liu R, Kong D, Wang S, Wang L, Yin B, Gao R. (2021) Attentive 3D-Ghost Module for Dynamic Hand Gesture Recognition with Positive Knowledge Transfer. Computational Intelligence and Neuroscience, 2021(5044916): pp 1–12. https://doi.org/10.1155/2021/5044916

[11] Attiah AZ, Khairullah EF. (2021) Eye-Blink Detection System for Virtual Keyboard. National Computing Colleges Conference (NCCC), pp 1-6. https://doi.org/10.1109/NCCC49330.2021.9428797.

[12] Shen Z, Elibol A, Chong NY. (2020) Understanding nonverbal communication cues of human personality traits in human-robot interaction. IEEE/CAA Journal of Automatica Sinica, 7(6): 1465-1477. https://doi.org/10.1109/JAS.2020.1003201

[13] Palinko O, Rea F, Sandini G, Sciutti A. (2016, October) Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration. International Conference on Intelligent Robots and Systems (IROS), pp. 5048-5054. https://doi.org/ 10.1109/IROS.2016.7759741

[14] Chandra B, Sharon HLU, Vignesh CP, Sriram R. (2020) Eye Blink Controlled Virtual Interface Using Opencv And Dlib. European Journal of Molecular & Clinical Medicine, 7(8), pp 2119–2126. https://ejmcm.com/article_4542.html

[15] Pasternak K, Wu Z, Visser U, Lisetti C. (2021) Let's be friends! A rapport-building 3D embodied conversational agent for the Human Support Robot. ArXiv preprint ArXiv:2103.04498. https://doi.org/10.48550/arXiv.2103.04498

[16] Saran A, Majumdar S, Short ES, Thomaz A, Niekum S. (2018, October 1) Human Gaze Following for Human-Robot Interaction. 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp 8615-8621. https://doi.org/10.1109/IROS.2018.8593580

[17] Saunderson S, Nejat G. (2019) How Robots Influence Humans: A Survey of Nonverbal Communication in Social Human–Robot Interaction. International Journal of Social Robotics, 11(4): 575–608. https://doi.org/10.1007/s12369-019-00523-0

[18] Li X. (2021) Design and implementation of human motion recognition information processing system based on LSTM recurrent neural network algorithm. Computational Intelligence and Neuroscience. https://doi.org/10.1155/2021/3669204

[19] Laghrissi FE, Douzi S, Douzi K, Hssina B. (2021) Intrusion detection systems using long short-term memory (LSTM). Journal of Big Data, 8(1): 1-16. https://doi.org/10.1186/s40537-021-00448-4