Design Studies and Intelligence Engineering
L.C. Jain et al. (Eds.)

doi:10.3233/FAIA220048

A HRI Framework Based on Eye Tracking

Xiaodong Zhao[a,1], Zheng Liu[b,c] and Shiwei Cheng[a]
[a] *Zhejiang University of Technology, School of Computer Science, Hangzhou 310023, China*
[b] *Zhejiang Provincial Key Laboratory of Integration of Healthy Smart Kitchen System, Ningbo 315336, China*
[c] *China Academy of Art, Hangzhou 310000, China*

Abstract. Existing human-robot interaction (HRI) technologies are not intuitive or effective enough for the disabled. Most of them can neither express their explicit intent (voice, posture, touch, etc.), nor communicate based on implicit intent. We propose a novel HRI framework for implicit intent reasoning based on eye tracking. The framework is designed for people with normal vision, and it will track and analyze their eye movements to infer their intents in a smart home environment. These intents are then sent to an assistive robot and guide the robot to accomplish specific tasks such as using keys to open the door. Two Experiments were carried out to validate the effectiveness of the framework. In the implicit intents classification task, the average classification accuracy of the two implicit intents (task-free visual browsing and task-oriented visual search) was 84.43% via the Support Vector Machine (SVM). In implicit intent reasoning task, Naive Bayesian (NB) networks were used to establish the intent knowledge base, reaching the highest accuracy of intent reasoning was 98%. These results proved that our framework could aid the elderly and the disabled to better adapt to the smart home environment in their daily life.

Keywords. Smart home, Eye tracking, Human-robot interaction, Implicit intent reasoning.

1. Introduction

In recent years, assistive robots are commonly used in smart home to help the elderly or disabled with mobility difficulties in their daily life [1]. The existing human robot interaction modes include voice [2], joystick [3], body contact [4], etc. However, for the speech or mobility-impaired individuals, extra efforts are needed to help them interact with these robots in a more effective and intuitive way. Efficient human-computer interaction is inseparable from human intent recognition. In the past decades, more and more researches on the modeling and recognition of human intent in psychology and cognitive science provide a new paradigm for the research of human computer interface (HCI) and HRI [5].

In general HCI, the user's explicit intent can be expressed by facial expression, text and gesture. However, their implicit intent appears to be vague, which is difficult for other individuals and the HCI systems to understand. Correctly recognizing the implicit intent of users is the key to develop an efficient HCI system. In recent years, in the

[1] Xiaodong Zhao, School of Computer Science, Zhejiang University of Technology, Hangzhou, People's Republic of China, E-mail: 1564666023@qq.com.

research of implicit intent recognition, researchers used electroencephalogram (EEG) [6], electrooculogram (EOG) [7] and electromyogram (EMG) [8] to analyze users' implicit intent and achieved good results.

In activities of daily living (ADL), eye fixation is another natural signal that can be used to infer a person's intent, which has yet to be fully studied. Eye movement are closely connected with human cognitive process [9], and directly related to the visual information in the scene. Eye movement, such as fixation, saccade and blink, contains several eye movement features: fixation duration, fixation times and so on. Studies have shown that the change of pupil size is related to cognitive process and visual information [10]. Human implicit intent can be classified into task-free visual browsing (TFVB) intent and task-oriented visual search (TOVS) intent [11]. TFVB refers to browsing some objects of interest without assigning a specific search task, TOVS refers to assigning an object search task to find objects related to the task.

The main contribution of this paper mainly includes the following aspects: 1) We propose a novel HRI framework to realize eye tracking-based implicit intent reasoning, which allows users to intuitively express their intent to the assistive robot by naturally watching objects. The experiment results showed that the framework designed in this paper is accurate and effective; 2) In user implicit intent recognition, TFVB and TOVS intent are classified and only when the TOVS intent is recognized will we move to the next step for intent reasoning. Experiment results have proved that the collected eye movement data shows great potential in classifying user’s intent; 3) During user intent reasoning process, an implicit intent knowledge base is established to reason the user's true intent by object combination.

2. Related work

Eye tracking is the process of measuring eye movement. The most concerned event of eye tracking research is to determine where humans or animals look such as fixation. Many previous studies have focused on fixation estimation on two-dimensional screen [12]. Fixation estimation based on two-dimensional screen has also been widely used in HCI research, and it is used to replace the function of mouse controlling the cursor [13]. The interaction between human and robot in HRI is different from the above-mentioned HCI mechanism, because robot has complex system, autonomy and cognition, and often functions in complex real-world environment [14].

At present, the research of HRI based on eye tracking mainly focuses on gazing at the specified position on the interface, so as to trigger the corresponding command to control the motion of the robot. For example, fixation is used to control the wheelchair to move [15]. Similarly, some researchers have tried to use eye tracking for teleoperation of mobile robots [16]. These studies are mainly focusing on some simple trigger commands, and do not involve the user implicit intent. The accuracy of correctly trigger commands depends on the precision of eye tracking.

Only a few studies use eye tracking to express users' high-level implicit intent, which involves multiple object detection in the real scene and the relationship between them [17]. However, the effectiveness, practicability and intuitiveness of HRI framework based on eye tracking still need to be verified.

Based on the above-mentioned researches, we propose an HRI framework for implicit intent reasoning based on eye tracking to understand user’s true intent and help the assistive robot better understand the user and serve the user.

3. Framework based on eye tracking

The framework can infer the implicit intent based on the user’s eye tracking data, combined with object detection technology to determine the object the user is looking at, and then uses the Naive Bayes algorithm. Judge the user's true intent, and finally combine with the assistive robot to help the user complete tasks.

3.1. Implicit intent judgment based on eye tracking

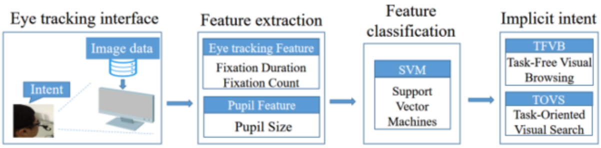


Figure 1. Framework for identifying user implicit intent.

Human eyes scan the surrounding environment and look at the objects in the scene intentally or unintentally, which are TOVS and TFVB respectively. TFVB is human's natural visual behavior, and TOVS is the behavior we want to detect. TOVS is a visual behavior oriented to task search. In order to detect human implicit intent, a TOVS and TFVB intent based on eye tracking and pupil feature changes is proposed (the framework is shown in Figure 1). Through the self-made head mounted eye tracker, record the eye tracking features and pupil features of the user viewing relevant scene images, including fixation duration, fixation times, pupil size and other features, extract these features, put these features into the Support Vector Machines (SVM) [18] classifier for feature classification, and finally output the implicit intent of the user.

3.2. Eye tracking based object detection

3.2.1. Fixation point clustering

The sampling speed of the original eye tracking fixation points collected by the eye tracking instrument is very fast, usually at 30Hz-50Hz, which can produce a large amount of eye tracking data in a short time. However, the duration of single eye tracking fixation point is very short, usually only 100ms. From the perspective of neuroscience, the human brain cannot process information in such a short time. Generally speaking, only when the fixation time is greater than 100ms, it can be considered that the fixation behavior can reflect the real activity of the human brain. In addition, when looking at the object, the human eye will shake unconsciously. In conclusion, it is necessary to cluster the original fixation points to effectively reduce the size of eye tracking data. The idea of clustering algorithm is to find a series of continuous fixed points whose time span is greater than a certain threshold (e.g. more than 100ms) and space is close, and take the average value of its coordinates as the coordinates of fixed points, the schematic diagram is shown in Figure 2.

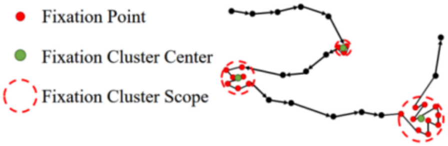


Figure 2. Fixation point clustering flow chart.

3.2.2. Object detection in scene image

The scene image is captured by the scene camera of the eye tracker after eye fixation clustering. Capture the scene image faced by the current user through the scene camera, obtain the object position and label in the current scene image through the object detection algorithm, and then calculate the object selected by the user through fixation in combination with the previous eye tracking fixation point data (The flow chart is shown in Figure 3). The target detection algorithm in this article uses the YOLOv3 model [19]. A single network structure is used to generate candidate regions while predicting the target category and location. Compared with Faster R-CNN [30], YOLOv3 produces a much smaller number of prediction frames, and each candidate area has only one true frame. These characteristics make the YOLOv3 algorithm faster and meet our needs for real-time detection of targets.

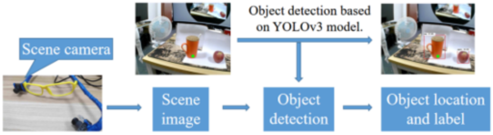

Scene camera
Object detection based on YOLOv3 model.
Scene image
Object detection
Object location and label

Figure 3. Flow chart of object detection in scene image.

3.3. Inference of human intent based on Naive Bayes

Through the previous work, we can judge the user's real-time fixation object, but only knowing a single fixation object can not well infer the user's real intent. We decompose the user' s single intent into multiple actions, and each action corresponds to an object, which forms a view of multiple objects associated with a user's intent. This correlation between the object watched through eye tracking and intent can be modeled as an intent prediction model, which is a Naive Bayesian (NB) classification model [20] , as shown in Figure 4. Objects are represented by O_i ($I = 1 \sim M$, where M is the total number of objects), and human intent users I_j ($j = 1 \sim N$, where N is the total number of possible intent types).

In the naive Bayesian classification model in this paper, the object set based on eye tracking detection is $O_i = \{O_1, O_2,..., O_M\}$, the user intent set is $I_j = \{I_1, I_2,..., I_N\}$, and the conditional probability of each recognized object for a certain intent is $P(I_j|O_i)$, The next step is to calculate the conditional probabilities $P(I_1|O_i)$, $P(I_2|O_i)$,..., $P(I_n|O_i)$ of each intent. Finally, by calculating the maximum probability, the intent corresponding to the maximum probability is the most likely user intent. If $P(I_k|O_i) = max\{P(I_1|O_i), P(I_2|O_i),..., P(I_n|O_i)\}$, according to the result, the current user intent is I_K.

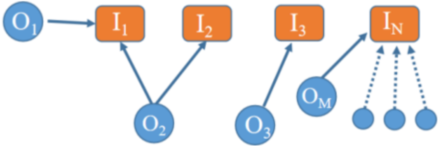


Figure 4. NB model for identifying the relationship between object and intent.

3.4. Human robot interaction with eye tracking fixation

The use of the assistive robot in indoor environment brings convenience to the elderly and the disabled with inconvenient movement. Therefore, it is necessary to consider designing a more intuitive interaction for users. As shown in Figure 5, this HRI framework based on eye tracking fixation includes four systems, eye tracking tracking system, user intent reasoning system, robot system and message agent system. Assistive robot is not only the visual display of the whole system effect, but also the part of direct interaction with people.

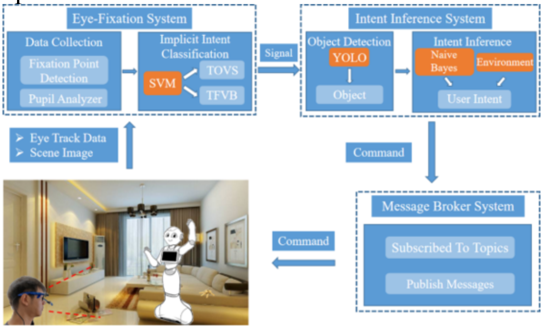


Figure 5. HRI system based on eye tracking fixation.

4. Experiments

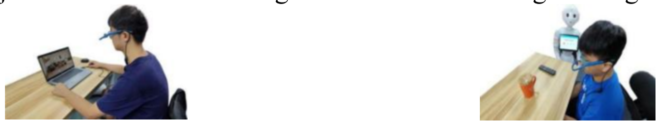
4.1. Participants

We recruited 10 healthy participants (7 males and 3 females, age 24.5 + 2.3). None of them was color blind or abnormal vision. This was the first time that all participants participated in the eye tracking experiment. All experiments were approved by the laboratory ethics committee of the author's University, and all participants agreed to participate before the beginning of the experiment.

4.2. Experimental equipment and scene

The experiment is divided into two parts, one is the detection of TFVB intent and TOVS intent, and the other is TOVS task judgment based on implicit intent knowledge base. The eye tracker uses the head mounted eye tracker developed in previous work [21], and the assistive robot uses the pepper humanoid robot of SoftBank[2].

In the first experiment, as shown in Figure 6 (a), the participants sat in front of the computer screen with an eye tracker on their head, completed different intent tasks through the prompts on the screen, and their eye tracking fixation positions and pupil sizes were recorded together. In the second experiment, as shown in Figure 6 (b), the participants also use the head mounted eye tracker to look at the objects in the real-time scene, the object that the user is looking at can be known through the algorithm.



(a)Experiment one

(b)Experiment two

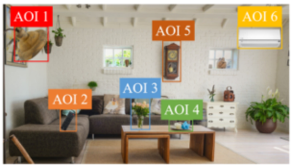
Figure 6. Experiment scenes.

[2] https://developer.softbankrobotics.com.cn

4.3. Implicit intent classification based on SVM

We use SVM to classify users' implicit intent, but eye tracking data needs to be trained before classification, so it is necessary to collect user related eye tracking experimental data. Participants were asked to sit in front of the computer screen and watch visual stimulation pictures. According to the prompts on the screen, they decided whether to perform TFVB intent task or TOVS intent task. Eye tracking features under different tasks are extracted and trained as training sets in SVM to obtain classification model data.

We propose a method to classify users' implicit intent using real-world images. Due to the page limit, we only showed the living room picture in Figure 7. It has six Area of Interest (AOI) according to the characteristics of its environment. Each AOI is a candidate region to prompt the experimenter to fixation. These AOI are the significant objects under the given visual stimulation conditions.



AOI 1: Light
AOI 2: Remote Control
AOI 3: Flower
AOI 4: Plate
AOI 5: Clock
AOI 6: Air Conditioner

Figure 7. AOI objects for visual stimulation images and settings.

During the experiment, the screen will first display instructions to the participants, and then the participants decide whether to execute TFVB intent task or TOVS intent task according to the instructions. The visual picture stimulation process is mainly divided into two steps:

- Step1) Visual browsing with no specific tasks: Participants browse stimulating pictures within a specified time to find their favorite content. Without specific task requirements, they browse stimulating pictures at will.
- Step2) Task oriented visual search: Participants first search for the specified object according to the screen prompt and the instructions on the screen.

All participants collected eye tracking data based on these two steps. As shown in Figure 8, the task prompt and stimulation image in the screen would be displayed for 5 seconds at a time, and there would be a 10 second rest between step 1 and step 2. Step 1 is executed only once in each round, while step 2 is executed six times according to the given AOIs, and the objects searched each time are different.

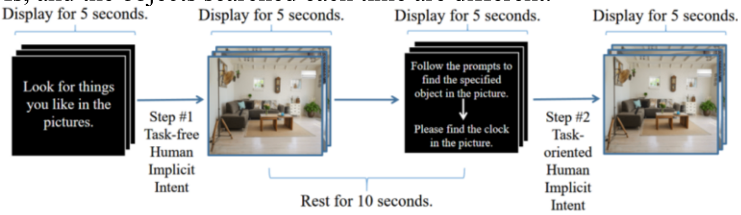

Display for 5 seconds.
Display for 5 seconds.
Display for 5 seconds.
Display for 5 seconds.
Look for things you like in the pictures.
Step #1 Task-free Human Implicit Intent
Follow the prompts to find the specified object in the picture.
Please find the clock in the picture.
Step #2 Task-oriented Human Implicit Intent
Rest for 10 seconds.

Figure 8. Experimental sequence of implicit intent events.

4.4. Establishment of implicit intent knowledge based on NB network

Through the analysis and investigation of indoor environment, this paper constructs a knowledge base of five potential TOVS intent tasks (as show in Table 1), including nine AOI potential areas (as show in Figure 9). Each AOI object may be the object that users want to watch. When the user looks at an AOI object through the eye tracker, the assistive robot will prompt the user and give a reasonable suggestion.

Table 1. The summary of dominant objects for each task

Tasks	Objects
Use the key to open the door	Room Key,Doorknob
Control the light with a switch	Light Switch,Lamp
Control the TV with the remote control	TV,Remote Control
Control the air conditioner with the remote control	Air Conditioning,Remote Control
Pour the wine into the glass	Glass,Bottle

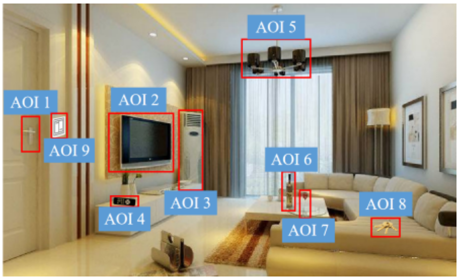


Figure 9. Indoor scene and corresponding AOI objects.

5. Result

5.1. Implicit intent classification based on SVM

TFVB and TOVS intent tasks showed different eye fixation characteristics. As shown in Figure 10, (a) and (c) are the heatmaps of eye tracking and fixation characteristics under TFVB intent task, and (b) and (d) are the heatmaps under TOVS intent task. During TOVS intent task, participants' fixation dwell time was longer than that of TFVB. In addition, it can be seen from the figure that the participants' fixation distribution is relatively concentrated when performing TOVS intent task, while the focus distribution in TFVB intent task is relatively scattered.

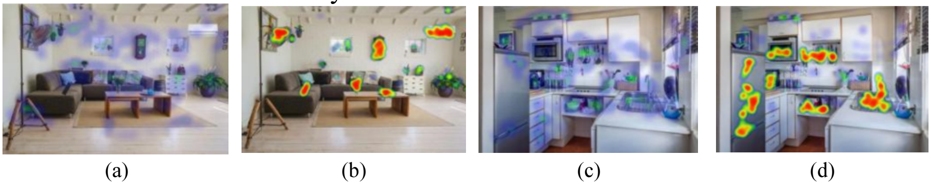
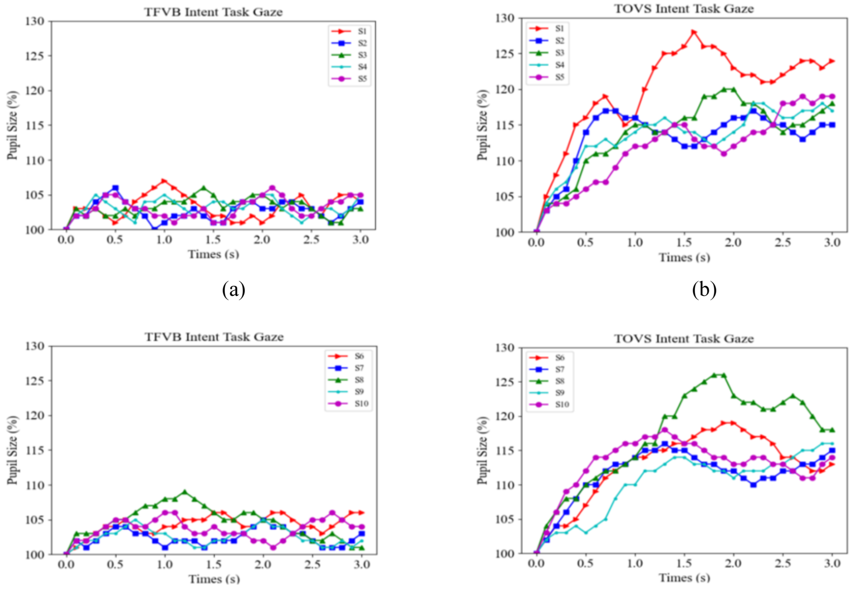


Figure 10. The heat map of eye tracking fixation feature.

The pupil size changes of participants under different tasks are shown in Figure 11. The 10 participants recruited were random divided into two groups. One group took the picture of the living room as the stimulus picture, and the other group took the picture of the kitchen as the stimulus picture. Both groups of participants had to complete TFVB and TOVS intent tasks. Figure 11 (a) (c) shows the pupil changes of participants under TFVB intent task, and (b) (d) shows the pupil changes of participants under TOVS intent task. As can be seen from the figure, the changes of pupil characteristics under the two tasks are obvious. The pupil size changes little under the TFVB intent task, but it increases significantly under the TOVS intent task.



(a) (b)

(c) (d)

Figure 11. Pupil size changes of participants under different tasks.

Through the above experiment, the eye tracking data sets of 10 participants were collected, including fixation duration, fixation times and pupil size. SVM was used to train and classify the data. In TFVB and TOVS intent tasks, training data sets account for 60% and test data sets account for 40%. Figure 12 shows the average detection accuracy of 10 participants under two tasks. The average detection accuracy of TFVB and TOVS intent tasks are 85.40% and 83.45%, respectively.

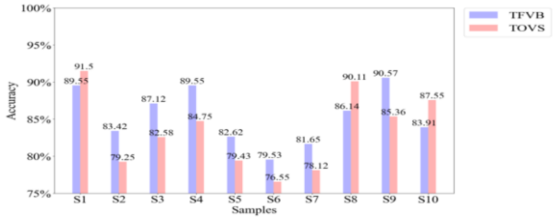


Figure 12. Task classification results based on SVM model.

5.2. Intent inference based on NB network

When the object identified by the object detection system meets the settings in the table, the system can infer the user's intent, and then send it to the robot, and the robot can know the task the user wants to perform. In this process, the accuracy of object detection and the object error of user search will cause intent inference errors, which will bring some errors. Figure 13 shows the results of the intent inference experiment participated by 10 participants who did not know the relevant intent and the corresponding main objects before participating in the experiment. The histogram shows the average detection accuracy of object detection and the accuracy of intent inference.

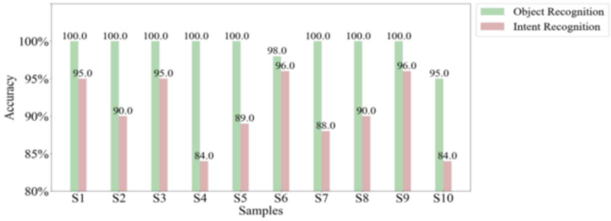


Figure 13. Results of intent inference based on Naive Bayesian model.

6. Discussion

The experiment results show that this HRI framework has great potential for use in the smart home environment. There are still many problems to be solved in this paper. For example, the scale of the implicit intent knowledge base is relatively small and the structure is simple; the object detection algorithm needs to be improved, and the system delay needs to be further reduced; there are fewer types of object detection, especially in the environment of smart home. In the future, we will consider expanding the knowledge base of implicit intent, and constructing a knowledge graph for users, increasing the richness, rationality, and applicability of the knowledge base, and making personalized recommendations for users. Consider the use of new algorithms for object detection, such as YOLOv5 [22], to shorten the object detection time.

7. Conclusions

We design a novel HRI framework of implicit intent reasoning based on eye tracking, which is used to infer users' intent in the environment of smart home, and then control the assistive robot to complete specific tasks. During the process of implicit intent judgment, experiments are designed and users' eye tracking data are collected, and SVM is used to classify users' TFVB and TOVS intent. Next, in the implicit intent reasoning process, the object detection algorithm is used to detect the object watched by the user in the real-time scene, and the implicit intent knowledge base is established by using Naive Bayesian network. This framework is to help the elderly and the disabled with mobility disabilities adapt to the environment of smart home and send their implicit intent to the assistive robot, so as to facilitate their daily life.

Acknowledgement

We thank all the volunteers, and the grant from Natural Science Foundation of China (No.62172368, 61772468), Fundamental Research Funds for the Provincial Universities of Zhejiang (No.RF-B2019001), and Zhejiang Provincial Key Laboratory of Integration of Healthy Smart Kitchen System (No.2020F04) .

References

[1] Mucchiani, C., Sharma, S., Johnson, M., Sefcik, J., Vivio, N., Huang, J., & Yim, M. (2017, September). Evaluating older adults' interaction with a mobile assistive robot. In 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 840-847). IEEE.

[2] Taylor, R. H., Menciassi, A., Fichtinger, G., Fiorini, P., & Dario, P. (2016). Medical robotics and computer-integrated surgery. Springer handbook of robotics, 1657-1684.

[3] Moreau, R., Pham, M. T., Tavakoli, M., Le, M. Q., & Redarce, T. (2012). Sliding-mode bilateral teleoperation control design for master–slave pneumatic servo systems. Control Engineering Practice, 20(6), 584-597.

[4] F. Zhang, A. Cully and Y. Demiris, "Probabilistic Real-Time User Posture Tracking for Personalized Robot-Assisted Dressing," in IEEE Transactions on Robotics, vol. 35, no. 4, pp. 873-888, Aug. 2019, doi: 10.1109/TRO.2019.2904461.

[5] Saponaro, G., Salvi, G., & Bernardino, A. (2013, May). Robot anticipation of human intentionintents through continuous gesture recognition. In 2013 International Conference on Collaboration Technologies and Systems (CTS) (pp. 218-225). IEEE.

[6] Zhao, M., Gao, H., Wang, W., & Qu, J. (2020). Research on human-computer interaction intentionintent recognition based on EEG and eye movement. IEEE Access, 8, 145824-145832.

[7] Huang, Q., He, S., Wang, Q., Gu, Z., Peng, N., Li, K., ... & Li, Y. (2017). An EOG-based human–machine interface for wheelchair control. IEEE Transactions on Biomedical Engineering, 65(9), 2023-2032.

[8] Bi, L., & Guan, C. (2019). A review on EMG-based motor intentionintent prediction of continuous human upper limb motion for human-robot collaboration. Biomedical Signal Processing and Con trol, 51, 113-127.

[9] Schütz, A. C., Braun, D. I., & Gegenfurtner, K. R. (2011). Eye movements and perception: A selective review. Journal of vision, 11(5), 9-9.

[10] Claudio, M.P., et al., The pupil dilation response to visual detection. Human Vision and Electronic Imaging XIII/SPIE-IS&T, 2008.p. 6806.

[11] Jang, Y. M., Mallipeddi, R., & Lee, M. (2014). Identification of human implicit visual search intentionintent based on eye movement and pupillary analysis. User Modeling and User-Adapted Interaction, 24(4), 315-344.

[12] Kar, A., & Corcoran, P. (2017). A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms. IEEE Access, 5, 16495-16519.

[13] Zhang, X., Liu, X., Yuan, S. M., & Lin, S. F. (2017). Eye tracking based control system for natural human-computer interaction. Computational intelligence and neuroscience, 2017.

[14] Komatsu, T., Kurosawa, R., & Yamada, S. (2012). How does the difference between users' expectations and perceptions about a robotic agent affect their behavior?. International Journal of Social Robotics, 4(2), 109-116.

[15] Shinde, S., Kumar, S., & Johri, P. (2018, September). A Review: Eye Tracking Interface with Embedded System & IOT. In 2018 International Conference on Computing, Power and Communication Technologies (GUCON) (pp. 791-795). IEEE.

[16] Carreto, C., Gêgo, D., & Figueiredo, L. (2018). An eye-gaze tracking system for teleoperation of a mobile robot. Journal of Information Systems Engineering & Management, 3(2), 16.

[17] Yu, Z., Kim, S., Mallipeddi, R., & Lee, M. (2015, July). Human intentionintent understanding based on object affordance and action classification. In 2015 International Joint Conference on Neural Networks (IJCNN) (pp. 1-6). IEEE.

[18] Awad, M., & Khanna, R. (2015). Support vector machines for classification. In Efficient Learning Machines (pp. 39-66). Apress, Berkeley, CA.

[19] Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

[20] Lowd, D., & Domingos, P. (2005, August). Naive Bayes models for probability estimation. In Proceedings of the 22nd international conference on Machine learning (pp. 529-536).

[21] Cheng, S. , Fan, J. , & Dey, A. K. . (2018). Smooth gaze: a framework for recovering tasks across devices using eye tracking. Personal and Ubiquitous Computing, 22(3), 1-13.

[22] Kuznetsova, A., Maleva, T., & Soloviev, V. (2020, December). Detecting apples in orchards using YOLOv3 and YOLOv5 in general and close-up images. In International Symposium on Neural Networks (pp. 233-243). Springer, Cham.