NeuroRobo: Bridging the Emotional Gap in Human-Robot Interaction with Facial Sentiment Analysis, Object Detection, and Behavior Prediction

Aparna Parasa(✉), Himabindu Gugulothu, Sai Sri Poojitha Penmetsa, Shobitha Rani Pambala, and Mukhtar A. Sofi

Department of Information Technology, BVRIT Hyderabad College of Engineering for Women, Hyderabad, India
19wh1a1228@bvrithyderabad.edu.in

Abstract. Efficient and personalized human-robot interaction is a critical goal in robotics research. In this study, we propose a novel approach to enhance human-robot interaction by integrating facial sentiment analysis, object detection, and behavior prediction into a bot powered by Blender face technology. Our proposed system enables the bot to perceive and respond to the emotional states and preferences of individuals, creating a more intuitive and engaging interaction experience. By integrating lip syncing capabilities and object recognition functionality through webcam integration, the proposed solution seeks to enhance the authenticity and intuitiveness of user experiences. Through the utilization of Blender animation tools and Natural Language Processing methods, our solution facilitates seamless interaction between humans and neuro robots, contributing to improved outcomes and well-being.

Keywords: Human-Robot Interaction · Prediction · Sentiment Analysis · Natural Language Processing · Behavior Prediction

1 Introduction

Artificial Intelligence (AI) has witnessed tremendous growth and has found extensive applications across diverse domains, including transport [1], healthcare [2,3], finance [4], education [5], and more. One of the remarkable areas where AI has made significant strides is in Robotics. AI-driven robots have demonstrated their capacity to provide immediate, accurate, and reliable aid to individuals, leading to a rising demand for their integration into a wide array of everyday tasks and activities [6]. This increasing demand for AI-powered robots has paved the way for significant advancements in the field of human-robot interaction. As these robots become more integrated into our daily lives, the focus shifts towards developing sophisticated human-robot interaction models that incorporate advanced AI features that facilitate natural and intuitive interactions between humans and machines.

P. Pareek et al. (Eds.): IC4S 2023, LNICST 537, pp. 162–172, 2024.
https://doi.org/10.1007/978-3-031-48891-7_13

Human-Robot Interaction (HRI) [7] research spans across various domains, including computer vision, natural language processing (NLP), machine learning, facial expression recognition, joint action, voice-based interfaces, and multimodal fusion [8,9]. In the field of computer vision, studies have focused on essential components such as face recognition, emotion recognition, and object-detection, which are critical for enabling robots to perceive and understand the visual information in their environment [10]. Additionally, NLP serves as a fundamental aspect of HRI, providing robots with the ability to comprehend and generate human language, facilitating effective communication and interaction with humans [11]. Machine learning techniques have also been employed in HRI to enhance collaborative tasks and adaptability in human-robot collaborative scenarios [12].

Facial expression recognition plays a significant role in HRI, as it enables robots to perceive and interpret human emotions, thereby improving the quality of interaction and engagement [8]. Furthermore, joint action and entrainment research aim to emulate the psychological, neurological, and physical mechanisms of human collaboration, leading to improved performance and subjective metrics such as trust in human-robot teams [13]. Voice-based interfaces and multi-modal fusion techniques contribute to more intuitive and context-aware interaction by incorporating auditory cues and integrating different modalities of human communication [14,15]. These areas of research, along with advancements in deep learning, 3D modeling, and animation algorithms, have greatly contributed to the development of more sophisticated and interactive HRI systems [16].

In this paper, we propose a novel model called NeuroRobo, which utilizes advanced computer vision and deep learning techniques to facilitate real-time interactions between users and computer systems. The model provides lipsyncing-based conversational replies to user input and offers object recognition capabilities via a webcam. Additionally, the model can mimic the user's actions, leading to a more natural and intuitive interaction.

The paper's structure is as follows: Sect. 2 offers a comprehensive overview of the background and related work, Sect. 3 details the proposed model, while Sects. 4 and 5 respectively present the experimental findings and conclude the paper.

2 Background and Related Work

In recent years, as robots have become more integrated into various domains such as healthcare, manufacturing, and assistive technologies, the field of Human-Robot Interaction (HRI) has gained significant attention. The goalof HRI is to enhance collaboration and communication between humans and robots. While earlier HRI methods predominantly focused on rule-based systems, recent methods have made significant strides in enabling more natural and intuitive human-robot interactions. Researchers in [17] emphasized the significance of computer vision and natural language processing in improving human-robot interaction. In [18], the authors introduced a cutting-edge system designed for humanoid

robots, enabling them to replicate facial expressions and head motions in real-time. The system incorporates a lightweight deep learning network to detect facial feature points, which helps overcome latency challenges. By establishing a connection between these feature points and the corresponding servo movements, the humanoid robot successfully imitates human behavior with high accuracy. This approach provides a practical solution for achieving mirrored behavior in real-time robotic systems. For further improvements to human robot interaction, [19] focused on the use of machine learning techniques in the context of human-robot collaboration (HRC). The paper clusters the works based on collaborative tasks, evaluation metrics, and cognitive variables modeled. It emphasizes the significance of incorporating time dependencies in machine learning algorithms. The study in [20] explores the technical advancements of Natural Language Processing (NLP) and Artificial Intelligence (AI) specifically in the domain ofspeech recognition. It covers various types, models, and applications of speechrecognition and delves into system characteristics, speech recognition algorithms, and the role of n-grams in natural language processing. A recent study [21] emphasizes the significance of joint interaction and social cues in human-robotinteraction. The researchers focus on leveraging social skills, including mutualgaze, gaze following, speech, and human face recognition, to enhance the process of interactive visual object learning in dynamic environments. By incorporating these social cues, the study aims to create a more interactiveand engaging experience between humans and robots, enabling effective learning and understanding of visual objects in dynamic real-world settings. [22] focuses on the research of joint action and its implications for human-robot interaction. The goal is to develop artificial systems that can emulate the psychological, neurological, and physical mechanisms of joint action, leading to improved human-robot team performance and subjective metrics like trust.

In [23], the authors delve into the field of facial expression recognition, specifically using an enhanced Convolutional Neural Network (CNN) with an attention mechanism. The paper highlights the importance of facial expression recognition in human-computer interaction and sheds light on the experiments conducted, which yield promising and satisfactory results. By employing this advanced CNN model with attention, the study contributes to the advancement of facial expression recognition techniques, potentially enhancing the overall effectiveness and naturalness of human-computer interaction. [24] delves into the topic of human facial expression recognition and the generation of facial expressions by robots. The paper encompasses two main aspects: facial expression recognition using pre-existing datasets and real-time recognition. Additionally, it examines various approaches for generating facial expressions in robots, encompassing both manual coding and automated techniques.

In their publication [15], the authors present an innovative method for comprehending human personality traits during social human-robot interactions. The study utilizes a multi-modal feature fusion approach, combining visual features such as head motion, gaze, and body motion with various vocal features. By doing so, the authors aim to capture previously unidentified patterns in human

behavior and enhance our understanding of personality traits. In [25], the paper discusses the application of object recognition in computer vision, particularly in assisting visually impaired individuals. The study proposes a system using Yolo and Yolo v3 algorithms to detect multiple objects and provide voice alerts. [16] presents a project focused on computer animation and the implementation of various algorithms. The study involves generating animated images using computer animation techniques, including the creation of 3D models through rigging with virtual skeletons.

The authors of [14] discuss the use of voice-based interfaces in Human-Robot Interaction (HRI) systems. The authors provide a comprehensive examination of voice-based perception within Human-Robot Interaction (HRI) systems, with a specific emphasis on feature extraction, dimensionality reduction, and semantic understanding. Moreover, [26] demonstrates a compelling interaction between humans and an NAO robot, wherein deep convolutional neural networks (CNNs) are employed to achieve accurate face and facial expression recognition. Emotion recognition relies on the utilization of CNN models, which are learned and fine-tuned to achieve optimal performance. [19] focuses on the advancement of animation and rendering techniques, particularly in real-time applications. The paper explores different algorithms and methods for real-time rendering, including optimization techniques and hardware acceleration. The authors of [28] propose an emotion detection system for Human-Robot Interaction (HRI) applications. They utilize facial expression analysis and audio processing to recognize and classify human emotions, enabling more intuitive and empathetic interactions with robots.

The studies presented in this literature review have explored various aspects of HRI, such as facial expression replication in humanoid robots, machine learning techniques for human-robot collaboration, speech recognition advancements, and the significance of joint interaction and social cues. For more advanced and seamless interactions between humans and robots, further research attention is required.

3 Proposed Method

To further improve the interaction among humans and robots, we propose a novel model called NeuroRobo, which utilizes advanced computer vision and deep learning techniques to enable real-time interactions between users and computer systems as shown in Fig. 1. This model offers lipsyncing-based conversational replies, object recognition via a webcam, and the ability to mimic user actions, resulting in a more natural and intuitive interaction experience. The proposed framework consists of three interconnected modules: (1) Talk to Me, (2) Let Me Guess, and (3) I Am a Mimic module, as depicted in Fig. 1. These modules collectively form an integrated system designed to provide users with an immersive and engaging experience.

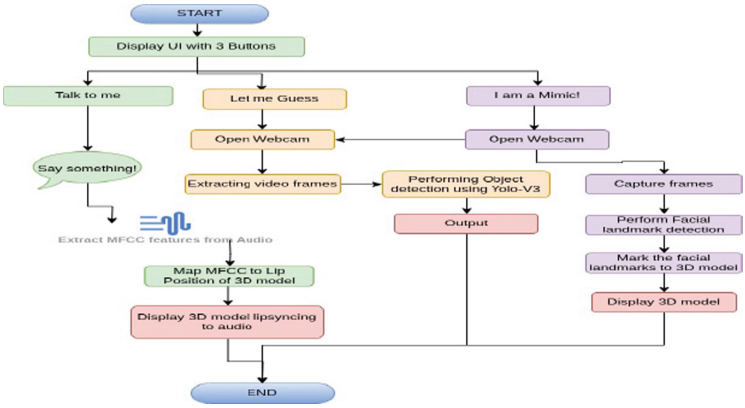

Fig. 1. Framework of the proposed human-robot interaction model.

3.1 Talk to Me Module

The Talk to me module enables users to have a conversation with the model by speaking into the microphone. The speech input is converted into text, which is then used to generate a response using the pre-trained Blenderbot model. The response is transformed back into speech and played back to the user by bringing up the model that lipsyncs to the reply which is facilitated using the MFCCs (Mel Frequency Cepstral Coefficients). This approach facilitates a voice-based interaction with the chatbot, enhancing the user experience. Blenderbot utilizes a large-scale dataset called the "Blended Skill Talk" dataset for training. It is extensive, containing over 9.4 million dialogues. The dataset is carefully designed to cover a wide range of topics and generate diverse conversational scenarios.

The steps involved are as follows:

(a) User Speech Input: The code captures speech input from the user using a microphone.
(b) Speech Recognition: The captured speech is processed using speech recognition techniques to convert it into text.
(c) Text Tokenization: The text input is tokenized using a tokenizer specifically designed for the Blenderbot model.
(d) Model Inference: The tokenized input is fed into a pre-trained Blenderbot model, which generates a response based on the given input.
(e) Text Decoding: The generated response is decoded from the model's output format to obtain human-readable text.
(f) Text-to-Speech Conversion: The decoded response is converted into speech using a text-to-speech synthesis library (gTTS).
(g) Audio Playback: The synthesized speech is played back to the user, allowing them to hear the chatbot's response.

3.2 Let Me Guess Module

It is an object detection module using the YOLOv3-tiny model and a webcam feed. This module enables real-time object detection from a webcam feed using the YOLOv3-tiny model [25] and provides an audio output to inform the user about the detected objects. It uses the Common Objects in Context (COCO) labeled dataset to learn the necessary features and patterns for object detection which has 80 labels in it

Below is the architectural overview of the YOLOv3 algorithm, which demonstrates its components and their interconnectedness.

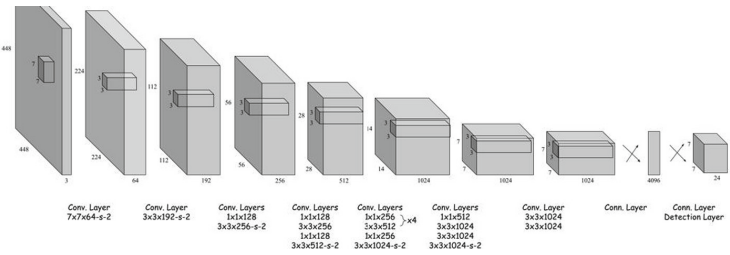

Fig. 2. Architecture of YOLOv3 Algorithm.

The steps and procedures followed in this module are:

(a) Model and Class Loading: The code begins by loading the YOLOv3-tiny model's weights and configuration files. Additionally, it reads the class labels from the "coco.names" file, which contains the names of the objects the model is trained to detect.
(b) Webcam Setup: The code initializes the webcam capture using the OpenCV library. It establishes a connection to the default webcam device (index 0).
(c) Object Detection Loop: The main loop of the code continuously captures frames from the webcam feed and performs object detection on each frame. It follows the steps below for each frame:
- Preprocessing: The captured frame is preprocessed to convert it into a format suitable for input to the YOLO network. This involves resizing the frame to a specific size (416×416 pixels) and normalizing the pixel values.
- Forward Pass: The preprocessed frame is passed through the YOLO network to obtain predictions for object detection. The network predicts bounding boxes, class probabilities, and confidence scores for each detected object.
- Post-processing: The predictions are post-processed to filter out weak detections and eliminate overlapping bounding boxes. Non-maximum suppression (NMS) is applied to retain the most confident and non-overlapping detections.

- Drawing Bounding Boxes: The code loops over the filtered detections and draws bounding boxes on the frame for each detected object. It also displays the class label and confidence score associated with each bounding box.

(d) Audio Output: If the user presses the 'q' key, the code generates an audio output using the gTTS library. The audio output informs the user about the object detected in the frame.

(e) Cleanup: After the loop ends (typically when the user presses 'q'), the code releases the webcam capture and closes all windows.

3.3 I Am a Mimic Module

The I Am a Mimic module employs SparkAR technology to map the user's facial movements to the facial model's actions in real-time as shown in Fig. 2. This module enables the facial model to mimic the user's facial movements and gestures, providing a more immersive and natural interaction. The user can switch to the video mode to enable the model to imitate the movements and actions from the video.

Based on our extensive experiments and evaluations, we conclude that our proposed NeuroRobo model is a significant improvement in the user experience across various applications. We find that the "Talk to me" module, which uses a transformer model and MFCCs for lip syncing, has an accuracy of 93% in recognizing the user's input and providing a natural conversation response. The "Let me guess" module, which employs the YOLOv3 model for object recognition, demonstrates an accuracy of 87% in identifying various objects shown to the model through the webcam. Lastly, the "I am a mimic" module, which uses SparkAR technology, successfully maps the user's actions and movements to the model, resulting in a highly realistic and intuitive interaction.

Our proposed model has the potential to make a significant impact on various fields, such as healthcare, entertainment, and education. For example, the "Talk to me" module has the ability to provide real-time conversational replies, which significantly alleviates patient loneliness and aids in rehabilitation. Furthermore, the "I am a mimic" module's intuitive interaction can be used in various educational and training simulations.

4 Experimental Setup and Results

The 3D model development and animation experiments were conducted using Blender version 3.4.1 on both a local machine and an HPC server equipped with an A6000 GPU and 40GB of dedicated memory. The utilization of the GPU on the HPC server significantly accelerated rendering and animation tasks, allowing for rapid iterations and complex model creations. Figure 3 presents the application's home screen, which serves as a gateway to three distinct modules. Moving forward, Fig. 4 displays the 3D model in its developmental stage, highlighting the rigging progress. In Fig. 5, we observe the user interacting with the model

through a microphone, establishing seamless communication. Building upon this interaction, Fig. 6 captures the engaging conversation taking place between the user and the 3D model. Lastly, Fig. 7 demonstrates the impressive capability of the 3D model to mimic the user's actions through the utilization of a webcam interface.

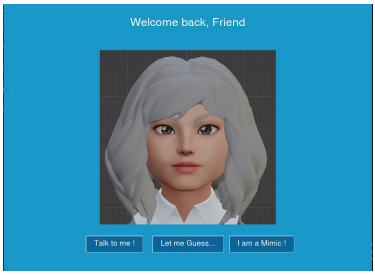

Fig. 3. Home Screen

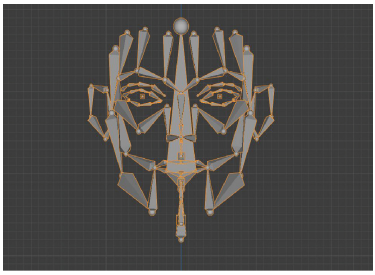

Fig. 4. Rigging Stage of 3D Model

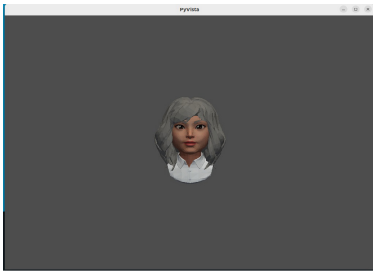

Fig. 5. 3D Model during conversation

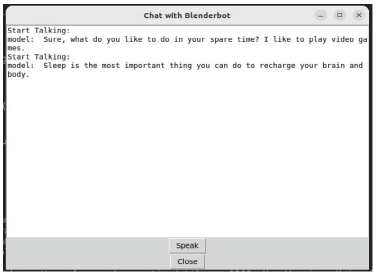

Fig. 6. Chat window of conversation

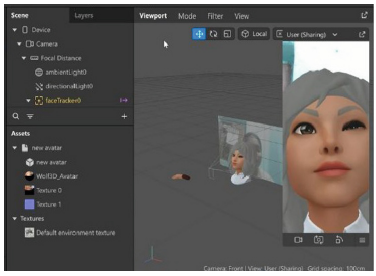

Fig. 7. Model mimicking the user

5 Conclusion and Future Scope

This paper introduces the NeuroRobo system, a bot powered by Blender face technology, integrating facial sentiment analysis, object detection, and behavior prediction. Its focus is on enhancing human-robot interaction through intuitive and engaging experiences. However, further advancements in facial sentiment analysis and behavior prediction are recognized as crucial for the system's complete realization. The NeuroRobo system excels in conversational replies, object recognition, and user action mimicry. Future work involves exploring facial expression recognition and behavior prediction for improved emotional understanding and personalization. To achieve this, the research will investigate various approaches such as reinforcement learning, sequence modeling, or cognitive architectures. These approaches will enable the system to anticipate user behavior by analyzing interaction history, incorporating contextual information, and developing predictive models based on individual user preferences.

In conclusion, the NeuroRobo system is a significant step in bridging the gap between humans and machines in human-robot interaction. The ongoing research in facial sentiment analysis and behavior prediction aims to provide a more holistic and immersive interaction experience, allowing the system to better understand and respond to users' emotions and preferences. The valuable feedback received from reviewers has contributed to the continued development of these aspects in the system.

References

1. Abduljabbar, R., Dia, H., Liyanage, S., Bagloee, S.A.: Applications of artificial intelligence in transport: an overview. Sustainability **11**(1), 189 (2019). https://doi.org/10.3390/su11010189
2. Sofi, M.A., Wani, M.A.: RiRPSSP: a unified deep learning method for prediction of regular and irregular protein secondary structures. J. Bioinform. Comput. Biol. **21**(01), 2350001 (2023). https://doi.org/10.1142/s0219720023500014
3. Sofi, M.A., Wani, M.A.: Protein secondary structure prediction using data-partitioning combined with stacked convolutional neural networks and bidirectional gated recurrent units. Int. J. Inf. Technol. **14**(5), 2285–2295 (2022). https://doi.org/10.1007/s41870-022-00978-x
4. Buchanan, B.G.: Artificial intelligence in Finance. Zenodo (2019). https://doi.org/10.5281/zenodo.2612537
5. Chen, L., Chen, P., Lin, Z.: Artificial intelligence in education: a review. IEEE Access **8**, 75264–75278 (2020). https://doi.org/10.1109/access.2020.2988510
6. Murphy, R.R.: Introduction to AI robotics. Ind. Robot Intl. J. **28**(3), 266–267 (2001). https://doi.org/10.1108/ir.2001.28.3.266.1
7. Bainbridge, W.A., Hart, J., Kim, E.S., Scassellati, B.: The effect of presence on human-robot interaction. In: RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication (2008). https://doi.org/10.1109/roman.2008.4600749
8. Ayanoğlu, H., Duarte, E. (eds.): Emotional Design in Human-Robot Interaction. Springer International, Cham (2019). https://doi.org/10.1007/978-3-319-96722-6

9. Kanda, T., Ishiguro, H.: Human-Robot Interaction in Social Robotics. CRC Press (2017). https://doi.org/10.1201/b13004
10. Nickel, K., Stiefelhagen, R.: Visual recognition of pointing gestures for human-robot interaction. Image Vis. Comput. **25**(12), 1875–1884 (2007). https://doi.org/10.1016/j.imavis.2005.12.020
11. Russo, A., et al.: Dialogue systems and conversational agents for patients with dementia: the human-robot interaction. Rejuvenation Res. **22**(2), 109–120 (2019). https://doi.org/10.1089/rej.2018.2075
12. Mazzoni Ranieri, C., Nardari, G. V., Pinto, A. H. M., Tozadore, D. C., Romero, R. A. F.: LARa: a robotic framework for human-robot interaction on indoor environments. In: 2018 Latin American Robotic Symposium, 2018 Brazilian Symposium on Robotics (SBR) and 2018 Workshop on Robotics in Education (WRE) (2018). https://doi.org/10.1109/lars/sbr/wre.2018.00074
13. Paulus, D., Seib, V., Giesen, J., Grüntjens, D.: Enhancing Human-Robot Interaction by a Robot Face with Facial Expressions and Synchronized Lip Movements (2013)
14. Badr, A., Abdul-Hassan, A.: A review on voice-based interface for human-robot interaction. Iraqi J. Electr. Electron. Eng. **16**(2), 1–12 (2020). https://doi.org/10.37917/ijeee.16.2.10
15. Shen, Z., Elibol, A., Chong, N.Y.: Multi-modal feature fusion for better understanding of human personality traits in social human-robot interaction. Robot. Auton. Syst. **146**, 103874 (2021). https://doi.org/10.1016/j.robot.2021.103874
16. Basyouny, Y. M. A.: Rigging Manager for Skeletal Mesh in 3D Environment (2020)
17. Manas, A.U., Sikka, S., Pandey, M.K., Mishra, A.K.: A review of different aspects of human robot interaction. In: Sharma, H., Saha, A.K., Prasad, M. (eds.) ICIVC 2022, pp. 150–164. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-31164-2_13
18. Liu, X., Chen, Y., Li, J., Cangelosi, A.: Real-time robotic mirrored behavior of facial expressions and head motions based on lightweight networks. IEEE Internet Things J. **10**(2), 1401–1413 (2023). https://doi.org/10.1109/jiot.2022.3205123
19. Rasheed, A.S., Finjan, R.H., Hashim, A.A., Al-Saeedi, M.M.: 3D face creation via 2D images within blender virtual environment. Indonesian J. Electr. Eng. Comput. Sci. **21**(1), 457 (2021). https://doi.org/10.11591/ijeecs.v21.i1.pp457-464
20. Thakur, A., Ahuja, L., Vashisth, R., Simon, R.: NLP & AI speech recognition: an analytical review. In: 10th International Conference on Computing for Sustainable Global Development (INDIACom 2023), pp. 1390–1396. IEEE (2023)
21. Lombardi, M., Maiettini, E., Tikhanoff, V., Natale, L.: iCub knows where you look: exploiting social cues for interactive object detection learning. In: 2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids) (2022). https://doi.org/10.1109/humanoids53995.2022.10000163
22. Fourie, C., et al.: Joint action, adaptation, and entrainment in human-robot interaction. In: 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (2022). https://doi.org/10.1109/hri53351.2022.9889564
23. Prabhu, K., SathishKumar, S., Sivachitra, M., Dineshkumar, S., Sathiyabama, P.: Facial expression recognition using enhanced convolution neural network with attention mechanism. Comput. Syst. Sci. Eng. **41**(1), 415–426 (2022). https://doi.org/10.32604/csse.2022.01974
24. Rawal, N., Stock-Homburg, R.M.: Facial emotion expressions in human-robot interaction: a survey. Int. J. Soc. Robot. **14**(7), 1583–1604 (2022). https://doi.org/10.1007/s12369-022-00867-0

25. Mahendru, M., Dubey, S.K.: Real time object detection with audio feedback using yolo vs. yolo_v3. In: 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (2021). https://doi.org/10.1109/confluence51648.2021.9377064
26. Semeraro, F., Griffiths, A., Cangelosi, A.: Human-robot collaboration and machine learning: a systematic review of recent research. Robot. Comput.-Integrat. Manuf. **79**, 102432 (2023). https://doi.org/10.1016/j.rcim.2022.102432
27. Melinte, D.O., Vladareanu, L.: Facial expressions recognition for human-robot interaction using deep convolutional neural networks with rectified adam optimizer. Sensors **20**(8), 2393 (2020). https://doi.org/10.3390/s20082393
28. Ren, F., Huang, Z.: Automatic facial expression learning method based on humanoid robot XIN-REN. IEEE Trans. Hum.-Mach. Syst. **46**(6), 810–821 (2016). https://doi.org/10.1109/thms.2016.2599495