Recognizing the Waving Gesture in the Interaction with a Social Robot

Conference Paper · October 2020

DOI: 10.1109/RO-MAN47096.2020.9223441

5 authors, including:

Giovanna Castellano
Università degli Studi di Bari Aldo Moro
241 PUBLICATIONS **1,959** CITATIONS
SEE PROFILE

Marco Cianciotta
Università degli Studi di Bari Aldo Moro
4 PUBLICATIONS **3** CITATIONS
SEE PROFILE

Berardina De Carolis
Università degli Studi di Bari Aldo Moro
162 PUBLICATIONS **1,767** CITATIONS
SEE PROFILE

Gennaro Vessio
Università degli Studi di Bari Aldo Moro
57 PUBLICATIONS **188** CITATIONS
SEE PROFILE

Recognizing the Waving Gesture in the Interaction with a Social Robot

Giovanna Castellano, *Member, IEEE*, Antonio Cervelione, Marco Cianciotta, Berardina De Carolis, *Member, IEEE* and Gennaro Vessio

***Abstract*— Humans use a wide range of non-verbal social signals while communicating with each other. Gestures are part of these signals and social robots should be able to recognize them for responding appropriately during a dialogue and being more socially believable. Gesture recognition is a hot topic in Computer Vision since a long time. This is particularly due to the fact that the segmentation of foreground objects from a cluttered background is a challenging problem, especially if it has to be performed in real-time. In this paper, we propose a vision-based framework for making social robots capable of recognizing and responding in real-time to a specific greeting gesture, namely the hand waving. The framework is based on a Convolutional Neural Network model trained to recognize hand gestures. Preliminary experiments in a lab setup with the social robot Pepper indicate that the robot correctly recognizes the wave gesture 90% of the times and answers appropriately in real-time by waving itself, thus increasing its social believability.**

I. Introduction

Social intelligence refers to the ability of robots to act socially with humans following behaviors and rules that are appropriate to their social role [1], [2]. When communicating, humans use a wide range of non-verbal social signals such as facial expressions, voice prosody, body postures and gestures. The analysis of non-verbal human behaviors during the interaction with social robots is crucial [3]. These signals should be recognized by the robot in order to respond appropriately during a dialogue and to exhibit more natural conversation capabilities with humans. Hence, processing social signals is important for having a correct perception of humans and their correct interpretation may result in an expression of a suitable social behavior by the robot, thus achieving a natural interaction with humans [4].

In this vein, Computer Vision may be of help in making social robots more aware of their users' characteristics, such as gender, age and emotions [5]. Another important task under the umbrella of Computer Vision concerns the recognition and interpretation of human gestures, particularly hand gestures. A gesture can be defined as any physical movement, either large or small, with which a person tries to communicate a specific message or instruction to other people [6]. There are several types of gesture: metaphoric gestures are those used to explain a particular concept; deictic gestures serve to point movements; iconic gestures are used to elaborate the meaning of the co-occurring speech; just to name a few. However, many gestures are culture-dependent and do not convey a unique meaning. In this paper, we focus our attention on a social gesture, namely waving, which is universally understood as a form of greeting. In fact, this type of gesture is very important to establish a communication among people independently, to a large extent, of their culture.

Fig. 1. The social robot Pepper.

The topic of gesture recognition has been widely investigated in the literature and lot of work has been done in recognizing gestures from images captured from traditional cameras (e.g., [7], [8]). Since social robots are spreading more and more into human life not only for entertainment, but also to assist users in their activities of daily living, or in teaching and educational settings, they should be capable of recognizing and responding to hand gestures. An example of application in which the recognition of gestures is of primary importance is to support migrants in recognising and learning the typical gestures of a culture and promoting their inclusion [9]. In particular, robots should be endowed with some communication skills so that users can interact with them just as they would intuitively do. They will need to be able to communicate naturally with people using both verbal and nonverbal signals. They will need to engage humans not only on a cognitive level, but on an emotional level as well. They will need a wide range of social-cognitive skills and a theory of other minds to understand human behavior, and to be intuitively understood by people [10].

The goal of this work is to equip a social robot with the ability of hand waving recognition using Computer Vision methods. In particular, as a testing platform we consider Pepper, a semi-humanoid robot developed by SoftBank Robotics (formerly Aldebaran Robotics),[1] equipped with hands and different sensors including a camera (Fig. 1) [11].

All the authors are with the Department of Computer Science, University of Bari "Aldo Moro", Bari, Italy.

[1] https://softbankrobotics.com

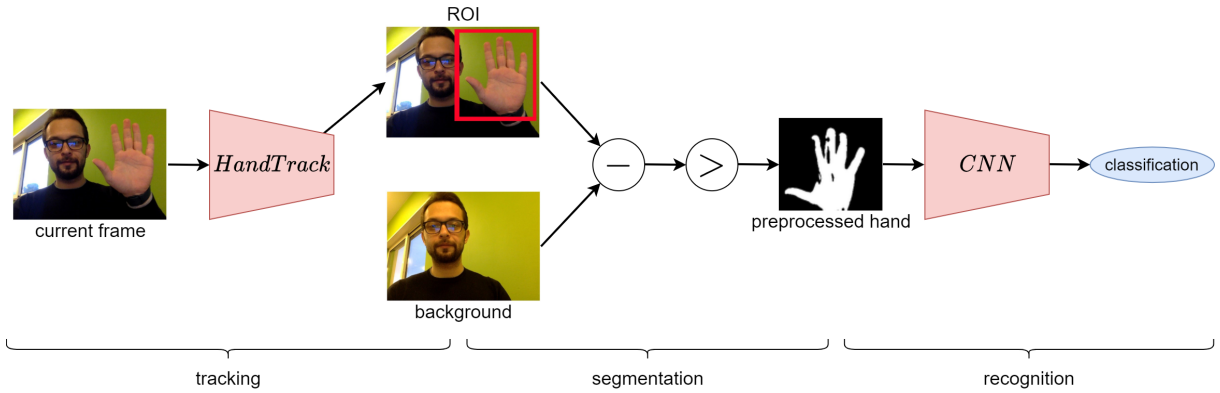

Fig. 2. Proposed workflow for waving gesture recognition.

To give Pepper the waving recognition ability, we have developed a framework based on a Convolutional Neural Network (CNN) that is trained to recognize the hand waving in video streams. CNNs are tailored to gesture recognition, thanks to their ability to approximate nonlinear relationships and to automatically learn meaningful representations from the low-level pixel features (e.g., [12], [13]).

The rest of the paper is organized as follows. Section II is about related work. Section III describes the proposed framework. Section IV provides experimental results. Finally, Section V concludes the paper.

II. Related Work

Gesture recognition is an active field of research in Computer Vision that has benefited from many pattern recognition and machine learning algorithms, such as support vector machines [14], temporal warping [15], [16], [17] and random forest classifiers [18] or other classification techniques [19]. In particular, several works have been proposed on hand gesture recognition for human-robot communication. In [20], a vision-based system is proposed to recognize specific gestures that have a communicative social meaning in a human-robot dialogue. In [21], a framework to define user gestures to control a robot is presented. In [22], communicative gestures are distinguished from daily living activities for an intuitive human-robot interaction.

More recently, deep learning models [23] have emerged as successful solutions for gesture recognition from videos. In particular, Convolutional Neural Networks have been widely applied. In [24], a 3D CNN in combination with spatio-temporal data augmentation is proposed for drivers' hand gesture recognition. In [25], a deep neural network architecture which takes advantage of a data fusion strategy is proposed for hand gesture recognition from videos. Deep neural networks are used in [26] to recognize gestures in real-time by considering only RGB information.

All the above methods are proposed as an off-line preliminary step towards developing human-robot interaction functionalities, hence they can work well only in simulated situations. To our knowledge, there are very few attempts to integrate the gesture recognition system into real social robots in order to achieve real-time gesture recognition. An example is given in [27], where a human-robot interaction system able to recognize gestures usually employed in human non-verbal communication is introduced. The system deals with dynamic gestures such as waving or nodding which are recognized using a dynamic time warping approach based on gesture-specific features computed from depth maps. Another example is given in [26], where a deep neural model is proposed to recognize dynamic gestures in an experimental set up using the humanoid robot Nimbro. In this work, we integrate a hand gesture recognition system in the humanoid robot Pepper, so that it can respond in real-time.

III. Proposed Method

The proposed framework is basically intended to provide the robot with the ability to (i) recognize the hand waving gestures and (ii) properly respond to them. In the following, we describe the workflow developed for hand gesture recognition, as well as the integration of the gesture recognizer into the Pepper robot.

A. Waving Gesture Recognition

We assume that the robot observes the scene by acquiring a video through its camera in the form of a sequence of video frames (30 fps), with each frame re-sized to 320×240 pixels. The robot has to detect the hand region from each frame and then track and analyze the moving hand to recognize the waving gesture. The workflow of the proposed method for gesture recognition is depicted in Fig. 2 and includes three main steps:

1) Hand ROI tracking;
2) Hand ROI segmentation;
3) Wave gesture recognition.

In the following, we describe in detail each step.

1) Hand ROI tracking: The first step is to detect and track the hand ROI through the frames. To do this, we use HandTrack, an end-to-end trainable CNN model that learns to track the human hand using a single RGB image in real-time [28]. HandTrack is based on Single Shot Detector (SSD) [29], which is a fast neural network model that can be easily integrated for designing user interactions (pointing,

Fig. 3. Sample images from the gesture dataset.

selections) across multiple platforms (web, mobile, desktop). In particular, we used a model pre-trained on the well-known MS COCO [30] and fine-tuned to the Egohands dataset [31]. This dataset contains high quality, pixel level annotations ($> 15,000$ ground truth labels) where hands are located across $4,800$ images. All images were captured from an egocentric view (Google glass) across 48 different environments (indoor, outdoor) and activities (card playing, puzzle solving, etc.).

2) Hand ROI segmentation: The second step for recognizing the waving gesture is to segment the hand region from each video frame so as to isolate the ROI corresponding to the hand. To do this, we create a background model by computing the running average of the background on 30 frames and then we calculate the absolute difference between the background model (updated over time) and the current bounding box including the hand. In this way, we obtain a difference image that holds the newly added foreground object, namely the hand. Then, we apply a thresholding to the difference image, so that pixels belonging to the hand region are white and pixels in all the other unwanted regions are black. Given the ROI corresponding to the hand, we apply an edge detector to find the contours of the hand.

3) Gesture recognition: Finally, the waving gesture is recognized using a standard Convolutional Neural Network that is trained on a dataset for gesture recognition. This dataset contains hand images related to three different gestures, namely fist, palm and swing (sample images are depicted in Fig. 3). Starting from this dataset, we created a training set of 1000 examples for each class and a test set of 100 examples for each class. We focused on these three gestures, as fist and swing can be easily confused with the waving gesture during the hand movement.

The proposed CNN consists of four convolutional layers, with an increasing number of filters (32, 64, 128 and 256), each with a stride of 2 and followed by a common ReLU activation function. Pairs of convolutional layers are interleaved by max pooling layers which downscale the output feature maps by a factor of 2. The last pooling layer is followed by a fully-connected layer with 2048 units and by a dropout layer (with dropout rate of 0.75), which is introduced to mitigate overfitting. The output layer is a classic *softmax* with a number of neurons equals to the number of classes. This model reflects a classic VGG-like architecture which is complex enough to handle our gesture dataset. The CNN was trained in a stochastic gradient descent fashion for 100 epochs with randomly selected mini-batches of size 64 and a learning rate of 0.001. To perform multi-class classification, the network was trained to minimize a categorical cross-entropy loss function.

B. Integration with Pepper

Pepper is a semi-humanoid robot developed to be used mainly as a commercial marketing tool and customer assistant in different application contexts such as stores, airports, hotels and so on. It has been used also in educational and playful contexts with children [32], [33].

The main difficulty one encounters when deploying a real-time application on Pepper is its limited computational capability and its lack of compatibility with new versions of programming languages such as Python. To overcome this limitation, we let all the computations needed for predicting the gestures be performed externally. To interact with the robot, we use NaoQi and Choregraphe, both developed by SoftBank Robotics. NaoQi is the name of the main software that runs on the robot and controls it. The framework is cross-platform and cross-language, with an identical API for both C++ and Python. Choregraphe, instead, is a multi-platform desktop application. It allows one to create animations and behaviors, test them on either a simulated robot or directly on a real one, monitor and control the robot. Since our gesture recognition system is developed in Python 3.7 while the above libraries are only available in Python 2.7, we developed a Web Service to integrate the recognition system into the Pepper robot. The Web service has been developed using Flask, that is a Web micro-structure written in Python.

The developed Web Service accepts POST requests attached with each frame acquired from the robot camera. In our case, the client side is represented by the Pepper robot. Initially, Pepper starts recording a video stream and each frame captured by the camera is sent to the Web server. Each frame is firstly transformed from a 3D array to a list and then inserted in a JSON file with the string 'photo' as a key. In the Web server, the data associated with the JSON 'photo' key are saved and converted to a 3D array. From the obtained frame, if the image contains a hand, the ROI is extracted and classified. Then, the prediction is returned to the client again in the form of JSON. If the frame has been classified as a waving gesture, then the robot takes actions accordingly. In particular, Pepper responds by greeting the user, both verbally and by performing the waving gesture itself.

IV. Experimental Results

The developed CNN module for waving gesture recognition was initially tested offline to evaluate its classification accuracy. The classification results in terms of precision, recall and F1 measure computed on the test set are shown in Table I. It can be seen that the model is able to achieve an average accuracy of more than 90% on the testing images, hence we can state that the model works almost perfectly in gesture recognition. An example of successful recognition is shown in Fig. 4. There are some problems only with the swing gesture. This may stem from the fact that the frame

Fig. 4. Waving gesture recognition.

TABLE I
CLASSIFICATION RESULTS.

Class	Precision	Recall	F1-score
Swing	0.83	1.00	0.90
Palm	1.00	0.79	0.88
Fist	1.00	1.00	1.00
Average	0.94	0.93	0.93

and the ROI extracted are taken in real-time, so the fingers are sometimes in positions not favorable for recognition.

After the integration of the CNN module in the Web Service (as described in III-B), the final system was tested with the Pepper robot in a laboratory setting. The experiment involved the recognition of the waving gesture performed by four subjects. Each subject performed the waving gesture while standing about 1.5m from the robot, as shown in Fig. 5. We observed that, under optimal conditions such as uniform background and proper amount of light, Pepper was able to correctly recognize the hand waving gesture in real-time and to answer both by voice and by replicating the wave gesture with its arm. However, we noticed that, due to the low resolution of the camera Pepper is equipped with (5 Megapixels), the presence of light reflecting objects (such as glasses), behind the user performing the gesture, may sometimes harm the recognition ability.

After evaluating the performance of the waving recognition module, another experiment was performed aiming at evaluating whether the social believability of the robot increases by adding the waving gesture recognition capability. Our hypothesis is that responsive behaviors produced by a robot appropriate to the social situation will induce in the human a perception of social intelligence. A concierge scenario was created in our Department entrance space in which the Pepper robot was welcoming people. Each person, who decided to participate in the experiment, was asked to enter in the Department reception and greet Pepper by waving in front of it. In total 30 people agreed to participate in the experiment and interacted with Pepper. They were young undergraduate students with an average age of 21 equally distributed by gender.

The system was tested in two conditions. In the first one, Pepper was not endowed with the waving gesture recognition capability and, then, when a person was waving at the robot, it did not greet back. In the second condition, Pepper was endowed with the waving recognition ability and, therefore, it greeted back to the recognized waving gesture. The participants were equally distributed in two groups, one

Fig. 5. Test cases in our laboratory. The Pepper robot is positioned in front of the human and recognizes the waving gesture continuously. After each recognition, Pepper gives a gestural feedback with the hand and says "hello".

for each condition. Thus 15 subjects were considered to test each condition. In both conditions, after the greetings, each participant was invited to evaluate (on a scale from 1 to 5) the perceived social believability during the greeting experience. This was achieved with the help of a facilitator who was present during the experimental sessions. The facilitator, besides counting how many times the robot was able to recognize the waving and greet back to the participants, was responsible for asking each participant to rate the experience by selecting a grade on a tablet in which the following question was displayed:

> *On a scale from 1 (not at all) to 5 (a lot), how much did you find believable Pepper's behavior in greeting you?*

We collected the ratings given by the students in the two situations and compared the results of the two groups using a t-test. The results are shown in Table II. It can be seen that in the first condition the rating given by subjects was lower than in the second condition, and the difference is significant according to the t-test p-value. This confirms that the robot is perceived as more believable when it is aware of the user gesture.

V. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a framework for waving recognition in human-robot interaction and we tested it with

the Pepper robot. We used two deep neural network models in a cascaded fashion, the first one aimed at identifying the hand ROI in a frame, while the second model was instructed to classify the gesture expressed within the previously obtained ROI. We integrated this classification workflow into a Web Service, in which the robot represented the client side. This solution was mainly due to the poor computational resources Pepper is equipped with. However, this solution has the potential to promote interoperability, as the proposed framework can be used to perform in parallel different deep neural network models conceived to solve different human-robot interaction tasks. Preliminary results in a real indoor environment showed that the developed waving recognition module is quite accurate and allows Pepper to recognize the gesture in real-time. Moreover, we performed an experiment to test whether the social capability of waving gesture recognition influenced the perception of the users interacting with Pepper in terms of social believability. We are aware that the study has been performed on a small number of subjects, however the experiment was carried out in a real setting and its results show that when the robot recognizes the person's greeting, it is perceived as more socially believable. We plan to perform another experiment in the wild with a larger number of subjects to better assess the effectiveness of our approach.

The developed framework is the first step toward a richer system for gesture recognition to support robust and natural interaction between a human and the social robot Pepper, enhancing multi-modal human-robot interaction. To this aim, further work is in progress to integrate other CNN modules for recognition of other hand gestures commonly used in communication. More in depth qualitative and quantitative evaluations of the proposed framework and its use with the Pepper robot are currently the topic of future work. Morevoer, it is worth noting that, in our study, the robot is equipped to have a "reactive" role when welcoming people (i.e., it responds to people's greeting); however, other studies (e.g., [34], [35] and [36]) focused on investigating the robot's perception as a "proactive" welcoming agent (i.e., the robot proactively greets people and draws their attention). As a future work, we want to explore ways to combine both approaches and estimate people's willingness to engage using the proposed platform.

Finally, it is worth noting that the Web Service-Pepper communication is not always efficient, because of possible network delay. This results in a reaction time not always short. As a future work, we also want to test the robot performance by using an embedded GPU, such as one of the well-known NVIDIA Jetson family, to be mounted directly on-board the robot.

TABLE II

COMPARISON OF BELIEVABILITY RATING IN THE TWO CONDITIONS.

	First condition (no waving recognition)	Second condition (waving recognition)
Mean	2.73	3.33
Std	1.20	0.92
p-value ($\alpha = 0.05$)	0.034	

REFERENCES

[1] A. H. Qureshi, Y. Nakamura, Y. Yoshikawa, and H. Ishiguro, "Robot gains social intelligence through multimodal deep reinforcement learning," in *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2016, pp. 745–751.

[2] T. J. Wiltshire, S. F. Warta, D. Barber, and S. M. Fiore, "Enabling robotic social intelligence by engineering human social-cognitive mechanisms," *Cognitive Systems Research*, vol. 43, pp. 190–207, 2017.

[3] H. Liu and L. Wang, "Gesture recognition for human-robot collaboration: A review," *International Journal of Industrial Ergonomics*, vol. 68, pp. 355–367, 2018.

[4] M. Moetesum and I. Siddiqi, "Socially believable robots," *Human-Robot Interaction: Theory and Application*, vol. 1, 2018.

[5] B. De Carolis, N. Macchiarulo, and G. Palestra, "Soft biometrics for social adaptive robots," in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 2019, pp. 687–699.

[6] M. A. Arbib, K. Liebal, S. Pika, M. C. Corballis, C. Knight, D. A. Leavens, D. Maestripieri, J. E. Tanner, M. A. Arbib, K. Liebal *et al.*, "Primate vocalization, gesture, and the evolution of human language," *Current Anthropology*, vol. 49, no. 6, pp. 1053–1076, 2008.

[7] A. Kanwal, M. U. G. Khan, K. A. Khan, M. S. Asif, S. M. Ahsan, and S. A. Raza, "Real time hand gesture recognition using PC camera," *International Journal of Computer Science and Information Security*, vol. 14, no. 11, p. 622, 2016.

[8] N. Lalithamani, "Gesture control using single camera for PC," *Procedia Computer Science*, vol. 78, pp. 146–152, 2016.

[9] B. D. Carolis, G. Palestra, C. D. Penna, M. Cianciotta, and A. Cervelione, "Social robots supporting the inclusion of unaccompanied migrant children: Teaching the meaning of culture-related gestures," *Journal of e-Learning and Knowledge Society*, vol. 15, no. 2, 2019.

[10] T. K. C. Breazeal, A. Takanishi, "Social robots that interact with people," *Springer Handbook of Robotics*, 2008.

[11] A. K. Pandey and R. Gelin, "A mass-produced sociable humanoid robot – Pepper: the first machine of its kind," *IEEE Robotics & Automation Magazine*, vol. 25, no. 3, pp. 40–48, 2018.

[12] G. Castellano, C. Castiello, C. Mencar, and G. Vessio, "Crowd detection for drone safe landing through fully-convolutional neural networks," in *International Conference on Current Trends in Theory and Practice of Informatics*. Springer, 2020, pp. 301–312.

[13] G. Castellano and G. Vessio, "Towards a tool for visual link retrieval and knowledge discovery in painting datasets," in *Italian Research Conference on Digital Libraries*. Springer, 2020, pp. 105–110.

[14] B. Liang and L. Zheng, "Multi-modal gesture recognition using skeletal joints and motion trail model," in *European Conference on Computer Vision*. Springer, 2014, pp. 623–638.

[15] A. Hernández-Vela, M. Á. Bautista, X. Perez-Sala, V. Ponce-López, S. Escalera, X. Baró, O. Pujol, and C. Angulo, "Probability-based dynamic time warping and bag-of-visual-and-depth-words for human gesture recognition in RGB-D," *Pattern Recognition Letters*, vol. 50, pp. 112–121, 2014.

[16] K. Kulkarni, G. Evangelidis, J. Cech, and R. Horaud, "Continuous action recognition based on sequence alignment," *International Journal of Computer Vision*, vol. 112, no. 1, pp. 90–114, 2015.

[17] M. Reyes, G. Dominguez, and S. Escalera, "Featureweighting in dynamic timewarping for gesture recognition in depth data," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, 2011, pp. 1182–1188.

[18] N. C. Camgöz, A. A. Kindiroglu, and L. Akarun, "Gesture recognition using template based random forest classifiers," in *European Conference on Computer Vision*. Springer, 2014, pp. 579–594.

[19] Z.-H. Chen, J.-T. Kim, J. Liang, J. Zhang, and Y.-B. Yuan, "Real-time hand gesture recognition using finger segmentation," *The Scientific World Journal*, vol. 2014, 2014.

[20] D. Michel, K. Papoutsakis, and A. Argyros, "Gesture recognition supporting the interaction of humans with socially assistive robots," in *Advances in Visual Computing*, ser. Lecture Notes in Computer Science, G. Bebis and et al., Eds. Springer International Publishing, 2014, vol. 8887, p. 793–804.

[21] M. Obaid, F. Kistler, M. Häring, R. Bühling, and E. André, "A framework for user-defined body gestures to control a humanoid

robot," *International Journal of Social Robotics*, vol. 6, no. 3, pp. 383–396, 2014.

[22] A. Chrungoo, S. Manimaran, and B. Ravindran, "Activity recognition for natural human robot interaction." in *Social Robotics*, ser. Lecture Notes in Computer Science, M.-A. W. M. Beetz, B. Johnston, Ed. Springer International Publishing, 2014, vol. 8755, p. 84–94.

[23] D. Wu, L. Pigou, P.-J. Kindermans, N. D.-H. Le, L. Shao, J. Dambre, and J.-M. Odobez, "Deep dynamic neural networks for multimodal gesture segmentation and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1583–1597, 2016.

[24] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3D convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 1–7.

[25] O. Kopuklu, N. Kose, and G. Rigoll, "Motion fused frames: Data level fusion strategy for hand gesture recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2103–2111.

[26] P. Barros, G. I. Parisi, D. Jirak, and S. Wermter, "Real-time gesture recognition using a humanoid robot with a deep neural architecture," in *2014 IEEE-RAS International Conference on Humanoid Robots*. IEEE, 2014, pp. 646–651.

[27] G. Canal, S. Escalera, and C. Angulo, "A real-time human-robot interaction system based on gestures for assistive scenarios," *Computer Vision and Image Understanding*, vol. 149, pp. 65–77, 2016.

[28] D. Victor, "HandTrack: A library for prototyping real-time hand tracking interfaces using convolutional neural networks," *GitHub repository*, 2017. [Online]. Available: https://github.com/victordibia/handtracking/tree/master/docs/handtrack.pdf

[29] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *European Conference on Computer Vision*. Springer, 2016, pp. 21–37.

[30] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.

[31] S. Bambach, S. Lee, D. J. Crandall, and C. Yu, "Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[32] F. Tanaka, K. Isshiki, F. Takahashi, M. Uekusa, R. Sei, and K. Hayashi, "Pepper learns together with children: Development of an educational application," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, 2015, pp. 270–275.

[33] G. Castellano, B. D. Carolis, N. Macchiarulo, and V. Rossano, "Learning waste recycling by playing with a social robot," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, 2019, pp. 3805–3810.

[34] P. Liu, D. F. Glas, T. Kanda, and H. Ishiguro, "Learning proactive behavior for interactive social robots," *Autonomous Robots*, vol. 42, no. 5, pp. 1067–1085, 2018.

[35] E. Saad, J. Broekens, M. A. Neerincx, and K. V. Hindriks, "Enthusiastic robots make better contact," in *IROS*, 2019, pp. 1094–1100.

[36] U. KC and J. Chodorowski, "A case study of adding proactivity in indoor social robots using belief–desire–intention (BDI) model," *Biomimetics*, vol. 4, no. 4, p. 74, 2019.