A robot-based surveillance system for recognising distress hand signal

Virginia Riego del Castillo[1,*], Lidia Sánchez-González[1], Miguel Á. González-Santamarta[1] and Francisco J. Rodríguez Lera[1]

[1]Department of Mechanical, Computer Science and Aerospace Engineering, Universidad de León, Campus de Vegazana s/n, León, 24071, Spain
*Corresponding E-mail: vriec@unileon.es
E-mails: vriec@unileon.es (Virginia Riego del Castillo); lidia.sanchez@unileon.es (Lidia Sánchez-González); mgons@unileon.es (Miguel Ángel González-Santamarta); fjrodl@unileon.es (Francisco Javier Rodríguez Lera)

Abstract

Unfortunately, there are still cases of domestic violence or situations where it is necessary to call for help without arousing the suspicion of the aggressor. In these situations, the help signal devised by the Canadian Women's Foundation has proven to be effective in reporting a risky situation. By displaying a sequence of hand signals, it is possible to report that help is needed. This work presents a vision-based system that detects this sequence and implements it in a social robot, so that it can automatically identify unwanted situations and alert the authorities. The gesture recognition pipeline presented in this work is integrated into a cognitive architecture used to generate behaviours in robots. In this way, the robot interacts with humans and is able to detect if a person is calling for help. In that case, the robot will act accordingly without alerting the aggressor. The proposed vision system uses the MediaPipe library to detect people in an image and locate the hands, from which it extracts a set of hand landmarks that identify which gesture is being made. By analysing the sequence of detected gestures, it can identify whether a person is performing the distress hand signal with an accuracy of 96.43%.

Keywords: computer vision, social robots, cognitive architecture, distress hand signal

1 Introduction

Robots are becoming more and more present in our daily lives. In particular, social robots are being deployed in several public environments, such as shopping malls [1], mental health centres to provide companionship to people [2] or educational places [3]. In these places, the robot could be used as a security tool that identifies safety or risky situations of people surrounding. The robot would recognise not only guns or fights but also perform biometric recognition [4] or other visual cues such as people shaking or choke gestures. It can be used indoors for surveillance in domestic areas, giving streaming of videos and other sensors [5]. On the other hand, outdoor security can be used for detecting people (passive infrared sensor) and checking if they are authorised (radio-frequency identification) [6].

A service robot that combines video surveillance and assistance to individuals in those risky situations can be of great interest to the community. Especially in certain high-risk scenarios where victims need help but cannot arouse the assailant's suspicions to ensure their safety, it is vital to identify the stealthy signals made by the victims to save their life. An inert, external agent, such as a robot, can be a watchdog that goes unnoticed by the aggressor and allows the victim's call for help to be identified.

For these risky situations, victims might display a hand signal created by the Canadian Women's Foundation [7], which is already internationally known for reporting kidnapping or domestic violence situations. For that matter, authors proposed a neural network that detects hand signs from images by identifying hand landmarks in a previous work [8]. In that paper, the considered images were captured with a laptop camera as if they were a video conference. By analysing the images, a certain sequence of hand signs is searched for in order to detect the SOS hand gesture [8].

This work adapts and integrates the gesture recognition pipeline into a robot's cognitive architecture as a robot skill together with other robot skills such as speech recognition, text to speech or navigation. Thus, the proposed skill enables the robot to recognise various kinesic signs of humans in real time, particularly, finding the SOS gesture. This would enhance the social capability of a service robot to help people in social environments, such as restaurants and shops. Besides, this research also details and publishes a dataset with images of different hand signs, including the signs of the distress gesture.

This paper is organised as follows. Section 2 describes the problem. Then, Section 3 presents the materials and methods employed in this work. Section 4 shows the obtained results, which are discussed in Section 5. Finally, Section 6 draws the main conclusions of the paper.

2 Description of the problem

A sign is defined as a movement done with your hands with a particular meaning. Otherwise, a signal or a gesture is the combination of signs to give a particular message. During the Covid-19 pandemic, the Canadian Women's Foundation (CWF) proposed a hand signal to secretly ask for help if women suffered domestic violence or required any kind of assistance. This distress hand signal consists of two signs presented in Figure 1. First, the hand is raised with the thumb tucked into the palm. Next, the thumb is trapped by your fingers.

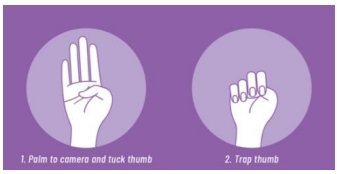


Figure 1: Signal for Help created by the Canadian Women's Foundation. Source: [7].

Hand gestures for human-computer interaction (HCI) have been developed with multiple technologies such as sensors in gloves or thermal, stereo and depth cameras [9]. A combination of colour and depth images with sEMG (Surface Electromyography) signals has also been used to train convolutional neural networks (CNN) that can recognise 10 different types of hand gestures with an accuracy of 92.45%.

There are many approaches to estimating hand pose using computer vision. Object's centre-of-mass and bounding box attributes extracted from a pre-processed image have been used to distinguish between four signs to control the motion of a robot [10]. In addition, the skin regions of the image are separated using the HSV colour space and then lines from the centre of the palm to each finger are calculated to classify the signs with a SVM [11]. However, neural networks, in particular CNNs, are the most widely used to solve this problem. In some cases, they are used to detect the hand in the image and identify which sign it shows [12]. In contrast, others only use them to recognise signs, detecting the hand with other methods such as HoG [13] or Haar cascade [14]. The RCE (restricted Coulomb energy) neural network is also used to segment hand images considering the skin colour and the sign recognition is determined by the centre-of-mass, which allows locating the position of the fingers [15]. Another commonly used solution is the use of the MediaPipe tool to calculate certain hand landmarks in order to recognise gestures in videos [16].

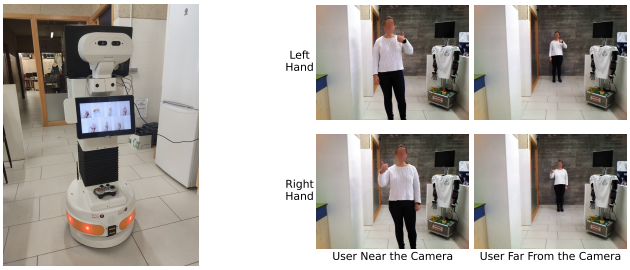
3 Materials and Methods

We have compiled a new dataset simulating real-life scenarios in order to develop our perception module to recognise the Signal for Help and integrate it into the robot's cognitive

architecture. The following sections describe the entire system.

3.1 Dataset

HaGrid [17] (HAnd Gesture Recognition Image Dataset) is a large dataset used for hand gesture recognition containing more than 500 thousand images and 18 different types of signs. However, it does not include any of the signs that constitute the Signal for Help. Therefore, a new dataset was created that includes images of the distress signal as well as other similar signals made by fourteen different users (8 males and 6 females). Images were acquired with the Asus camera integrated into a TIAGo robot (Figure 2a). Each user[1] made seven different signs with both hands (see Figure 3), both far away and close to the camera (2 and 4 meters, respectively). So, a greater variety of images was obtained. The frames of the acquired videos were manually filtered to reject those images where the user did not make any signal.



(a) TIAGo robot. (b) User recorded by the robot.

Figure 2: Acquisition system.

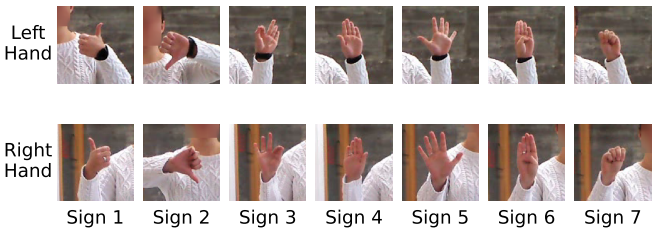


Figure 3: Signs done by one of the users with the left hand (first row) and right hand (second row). Signal for Help is formed by Sign 6 followed by Sign 7.

The sequence Sign 6 - Sign 7 represents the Signal for Help. As Figure 3 shows, there is a slight difference between Sign 7 and Signs 1 and 2, as only the position of the thumb varies. In addition, Sign 6 only differs from Signs 4 and 5 in the distance between the fingers.

[1] All users agreed to be recorded for scientific purposes according to current data protection regulations.

Figure 4 displays the distribution of the considered dataset by sign and hand. This ALMOST (hAnd LandMarks fOr Sos gesTure) dataset is publicly available in [18].

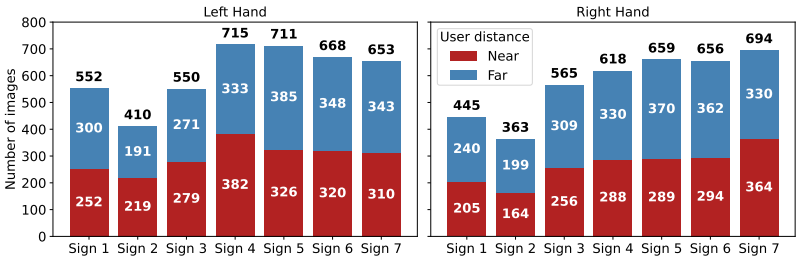


Figure 4: Frame distribution of the dataset. Users represent left and right hand signs both near and far from the camera.

For the experiments, four users were randomly selected for testing (two males and two females) and the rest for training. Figure 5 shows how the dataset is divided into training and testing.

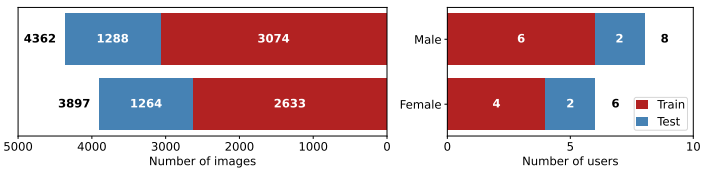


Figure 5: Distribution of the training and test data by number of images (left side) and number of users (right side).

3.2 Cognitive Architecture

Cognitive architectures are used to produce the behaviours of these robots, generating the social skills that robots need to interact with humans. There are three types of cognitive architectures [19]: symbolic, emergent and hybrid. The most extended type is the hybrid architecture which uses a deliberative system in its core and employs several emergent systems to create the robot skills. Moreover, facing social interaction between humans and robots is a complex task that has been boarded using different mechanisms. There are two types of Human-Robot Interaction (HRI) used in cognitive architectures [20]: verbal, which is the most common way of interaction between robots and humans; and non-verbal.

The cognitive architecture used in this research is MERLIN [21], a hybrid cognitive architecture whose aim is to produce behaviours in robots. MERLIN is composed of two main systems: deliberative and behavioural. Each system is divided into two layers.

The deliberative system includes the Mission Layer and the Planning Layer. The Mission Layer produces high-level goals for the robot. The Planning Layer has to generate

plans using the knowledge of the robot. A plan is a sequence of actions that can achieve the robot's goals. This layer also has a knowledge base that stores the knowledge about the robot's environment.

The behavioural system comprises the Executive Layer and the Reactive Layer. The Executive Layer involves the actions that the robot can perform. These actions can be implemented as state machines. The Reactive Layer gathers the skills of the robot, such as navigation, text-to-speech, speech-to-text and perception.

The gesture recognition system presented in this work is a skill of the robot that can be found in the perception module of the Reactive Layer. On one hand, it can be used to update the knowledge that the robot has about the person in the environment. As a result, that knowledge can be used to produce more complex plans. On the other hand, the signs and the gestures detected can be used in the actions of the Executive Layer.

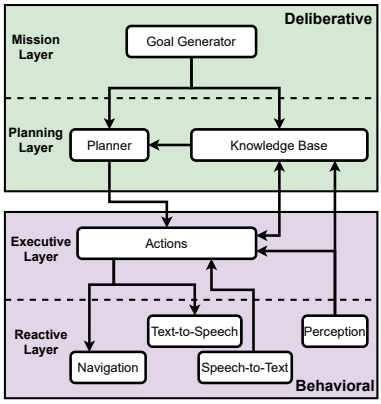


Figure 6: MERLIN cognitive architecture.

3.3 Perception Module

This module allows the robot to recognise hand gestures made by people by analysing the videos acquired with its camera. The following subsections describe the complete process, which is shown in Figure 7.

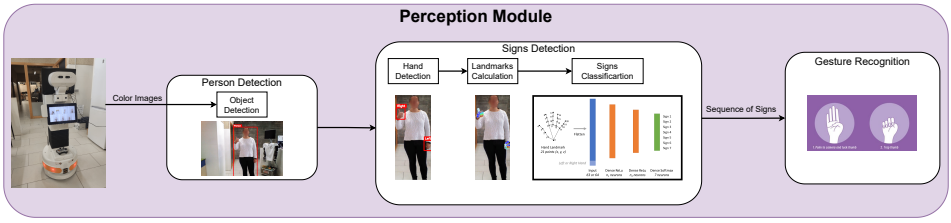


Figure 7: Perception module pipeline. Hand landmarks are extracted and classified from each image. Then frames of the video are processed to obtain the sequence of signs.

3.3.1 Hand landmark calculation

MediaPipe [22] is a framework that allows real-time hand tracking in colour images [23] by locating the bounding box of the hand and calculating certain landmarks of the hand to predict its skeleton.

However, as larger distances from the user to the camera affect hand detection performance, a model such as YOLOv5 [24] (specifically the smaller version, YOLOv5s), has been used to detect the person in the image and zoom in to where their hands are. These regions defined by a bounding box are cropped and an identification of the hand side is carried out. This hand-only image is then passed to MediaPipe to obtain the hand landmarks. Figure 8 displays an example of the complete pipeline to get hand landmarks from a frame.

Figure 8: Pipeline of the process to get hand landmarks. From the original image (a), people are detected (b) and then the image is zoomed in the bounding box region (c). Finally, MediaPipe is used to detect the hand (d) and get hand landmarks (e).

Hand landmarks are represented by 21 points of three dimensions (x, y, z), as can be seen in Figure 9. These points correspond to the pixel positions of the image. So each point is normalised to a position of the bounding box of the hand and varies from 0 to 1.

3.3.2 Sign classification

The perception module includes a neural network architecture designed to classify signs. The input of the proposed model is the 63-feature vector (21 points $\times$ 3 dimensions) plus the hand side (left or right). Otherwise, the architecture has two dense hidden layers of n_1 and n_2 neurons with ReLU activation. Finally, the output layer if formed by 7 neurons with softmax activation, giving the probability of being one of the considered signs. Figure 9 displays a scheme of the proposed network architecture.

3.3.3 Sign sequence processing

Previous sections detailed how each frame of a video is processed to calculate a set of hand landmarks and classify them to identify the represented sign. Therefore, this procedure can be applied to a complete video (set of frames). First of all, frames of the video are processed to identify which sign is being shown and its probability. Those signs that are

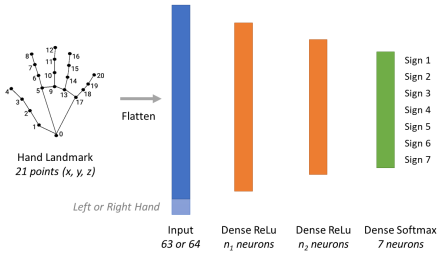


Figure 9: Scheme of the network architecture composed by two hidden layers of n_1 and n_2 neurons. Input is a vector of 64 values (21 values from the 3D hand landmarks points and 1 from the hand side).

detected with less than 70% of confidence are rejected because they corresponds with no signs or those frames where the hand is changing the sign. So, the remaining detected signs are considered as the actual sign identification of the system. After that, a certain sequence of signs that means the Signal for Help (Sign 6 and Sign 7) are determined. The completed process can be analysed in Figure 10. This process can be done in recorded video or in a real-time system.

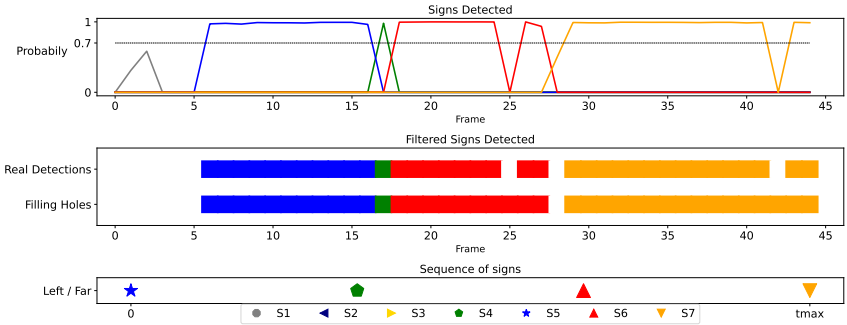


Figure 10: Process to determine the sign sequence. Each frame of the video is processed identifying a certain sign with a confidence probability (first row). Detections with less than 70% of confidence are rejected and spaces between the same signs are filled (second row). Finally, a sign sequence is extracted (third row).

4 Experimental results

For experiments, the proposed architecture explained in Section 3.3.2 is configured with different number of neurons in the first n_1 a second layer n_2. One is a smaller network N_{20-10} with $n_1 = 20$ and $n_2 = 10$. The other is a bigger one N_{30-15} with $n_1 = 30$ and $n_2 = 15$. In addition, two different inputs were taken into account, one formed by just the hand landmarks and other configuration N^{side} combining the landmarks with information about the hand side (left or right). All these configurations were trained with an Adam

optimizer and an exponential learning rate, which is initially settled up to $5 * 10^{-3}$ and a decay rate of 0.96 for every 25 steps. Categorical cross-entropy loss during 100 epochs with a batch size of 128 was also established. As the considered sign classes are not completely balanced (see Figure 4), classes are weighted during the training considering the number of images per class.

Experiments show the best results are achieved with the network with more neurons; although that implies more trainable parameters, the inference time keeps the same (31.29 milliseconds per input). Table 1 gathers the obtained results. As it can be observed, the best training results are obtained by the N^{side}_{30-15} network and the left-right information is not required as there is a slight improvement, with an accuracy of 97.3% for the training set and 92.8% for the test set.

Table 1: Accuracy achieved by the different configurations of the model proposed to identify 7 different hand signs.

Model	Trainable parameters	Inference time (ms)	Training accuracy	Test accuracy
N_{20-10}	1567	31.29	0.9686	0.9197
N^{side}_{20-10}	1587	31.29	0.9634	0.9224
N_{30-15}	2497	31.29	0.9727	**0.9275**
N^{side}_{30-15}	2527	31.29	**0.9737**	0.9248

The N^{side}_{30-15} network has been chosen for the final perception module. Then, users were asked to reproduce the SOS gesture again with both hands (left and right) and at different distances to the camera (near and far). The perception module was activated to get the sign sequence for a video and the Signal for Help was represented as a the Sign 6 followed by Sign 7. The combination with Sign 5 in between is also admitted due to its similarity to Sign 6. As it can be appreciated in Figure 11, only in two of the fifty-six videos the gesture was not detected, which is a 96.43% of accuracy in the detection of the distress gesture. All code and experiments are available in [25].

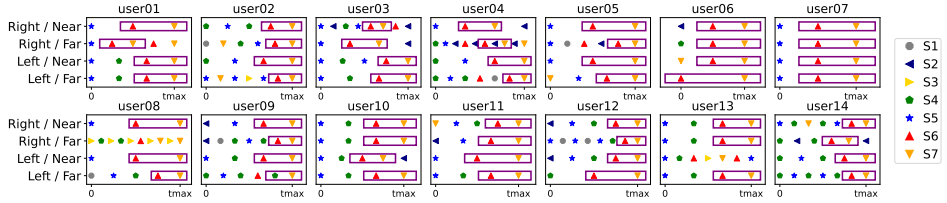


Figure 11: Users sequence of signs done with both hands (left/right) and distances to the camera (near/far). Detected SOS gestures are highlighted with a rectangle.

5 Discussion

Comparing the obtained results with the existing proposals that classify hand signs using neural networks, we can analyse the performance of the proposed perception module (see Table 2). In [26], a CNN with an image input of 5 dimensions is used, getting a 92.45% of accuracy, which is a little bit less than our approach but our architecture is simpler so faster for training. Other researchers also use the hand landmarks provided by MediaPipe, getting 87.5% of accuracy using landmarks and features extracted from the distance between points during the time [27]; our method outperforms these results as well as considers a smaller number of inputs in the employed architecture. In previous research [8], the proposed network achieved a 79.5% of accuracy for sign classification and a 75.67% for SOS gesture detection. These results are improved with the Perception Module proposed in this paper.

Table 2: Comparison of state art results for hand sign classification

Method	Signs	Network Input	Detection Accuracy	SOS gesture Detection
[26]	10	RGB-D image with sEMG $(160x160x5)$	92.45%	-
[27]	13	Mediapipe Hand Landmarks and features during time (882)	87.5%	-
[8]	7	Mediapipe Hand Landmarks $(21x3)$	79.5%	75.67%
Perception Module	7	Mediapipe Hand Landmarks with hand identification $(21x3+1)$	**92.48%**	**96.43%**

6 Conclusions

Domestic and gender-based violence is still a problem in our society. For this reason, the Canadian Women's Foundation proposed a hand gesture to be able to call for help in case of danger. The presented approach improves the cognitive system of a social robot by including a perception module that detects the Sign For Help by means of a vision-based system. Thus, a risky situation can be identified without alerting the aggressor. For this purpose, the ALMOST (hAnd LandMarks fOr Sos gesTure) dataset has been published with images of 14 users making 7 different signs with both hands (left and right) and at different distances to the camera (near and far). The set of signs includes the distress signal signs and similar signs. For this dataset, the proposed model recognises the signs with an accuracy of 92.48% in the test set and the sequence of signs representing the distress hand signal with 96.43% accuracy, which satisfies the expected performance.

Funding

EDMAR Project PID2021-126592OB-C21 funded by MCIN/AEI/10.13039/501100011033 and by ERDF A way of making Europe.

Acknowledgements

Virginia Riego acknowledges Universidad de León for its funding support for her doctoral studies. Miguel Á. González-Santamarta acknowledges an FPU fellowship provided by the Spanish Ministry of Universities (FPU21/01438). We also express our grateful to Centro de Supercomputación de Castilla y León (SCAYLE) for its infrastructure support.

References

[1] M. Niemelä, P. Heikkilä, H. Lammi, and V. Oksman, "A social robot in a shopping mall: Studies on acceptance and stakeholder expectations," *Social robots: Technological, societal and ethical aspects of human-robot interaction*, pp. 119–144, 2019.

[2] A. A. Scoglio, E. D. Reilly, J. A. Gorman, and C. E. Drebing, "Use of Social Robots in Mental Health and Well-Being Research: Systematic Review," *Journal of Medical Internet Research*, vol. 21, no. 7, e13322, 2019. DOI: 10.2196/13322.

[3] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka, "Social robots for education: A review," en, *Science Robotics*, vol. 3, no. 21, eaat5954, Aug. 2018, ISSN: 2470-9476. DOI: 10.1126/scirobotics.aat5954.

[4] C. Álvarez-Aparicio, Á. Guerrero-Higueras, M. González-Santamarta, A. Campazas-Vega, V. Matellán, and C. Fernández-Llamas, "Biometric recognition through gait analysis," *Scientific Reports*, vol. 12, no. 1, 2022. DOI: 10.1038/s41598-022-18806-4.

[5] G. Song, K. Yin, Y. Zhou, and X. Cheng, "A surveillance robot with hopping capabilities for home security," *IEEE Transactions on Consumer Electronics*, vol. 55, no. 4, pp. 2034–2039, 2009. DOI: 10.1109/TCE.2009.5373766.

[6] S. Meghana, T. V. Nikhil, R. Murali, S. Sanjana, R. Vidhya, and K. J. Mohammed, "Design and implementation of surveillance robot for outdoor security," in *2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, 2017, pp. 1679–1682. DOI: 10.1109/RTEICT.2017.8256885.

[7] J. Howard, *Signal For Help — Use Signal to Ask for Help*, en-CA. [Online]. Available: https://canadianwomen.org/signal-for-help/.

[8] R. Viejo-López, V. Riego del Castillo, and L. Sánchez-González, "Hand SOS gesture detection by computer vision," in *15th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2022) Proceedings*, Springer, 2022, pp. 22–29.

[9] M. Oudah, A. Al-Naji, and J. Chahl, "Hand Gesture Recognition Based on Computer Vision: A Review of Techniques," *Journal of Imaging*, vol. 6, no. 8, p. 73, Jul. 2020, ISSN: 2313-433X. DOI: 10.3390/jimaging6080073.

[10] A. M. Faudzi, M. H. K. Ali, M. A. Azman, and Z. H. Ismail, "Real-time Hand Gestures System for Mobile Robots Control," *Procedia Engineering*, International Symposium on Robotics and Intelligent Sensors (IRIS 2012), vol. 41, pp. 798–804, 2012. DOI: 10.1016/j.proeng.2012.07.246.

[11] R. C. Luo and Y. C. Wu, "Hand gesture recognition for Human-Robot Interaction for service robot," in *2012 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, Sep. 2012, pp. 318–323. DOI: 10.1109/MFI.2012.6343059.

[12] O. Köpüklü, A. Gunduz, N. Kose, and G. Rigoll, *Real-time Hand Gesture Detection and Classification Using Convolutional Neural Networks*, arXiv:1901.10323 [cs], Oct. 2019. DOI: 10.48550/arXiv.1901.10323.

[13] P. N. Huu and T. Phung Ngoc, "Hand Gesture Recognition Algorithm Using SVM and HOG Model for Control of Robotic System," *Journal of Robotics*, vol. 2021, e3986497, 2021. DOI: 10.1155/2021/3986497.

[14] D. N. Fernández, "Development of a hand pose recognition system on an embedded computer using Artificial Intelligence," in *IEEE XXVI International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, 2019, pp. 1–4. DOI: 10.1109/INTERCON.2019.8853573.

[15] X. Yin and M. Xie, "Finger identification and hand posture recognition for human–robot interaction," *Image and Vision Computing*, vol. 25, no. 8, pp. 1291–1300, 2007. DOI: 10.1016/j.imavis.2006.08.003.

[16] G. Benitez-Garcia, J. Olivares-Mercado, G. Sanchez-Perez, and K. Yanai, "Ipn hand: A video dataset and benchmark for real-time continuous hand gesture recognition," in *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, 2021, pp. 4340–4347.

[17] A. Kapitanov, A. Makhlyarchuk, and K. Kvanchiani, *HaGRID - HAnd Gesture Recognition Image Dataset*, arXiv:2206.08219 [cs], Jun. 2022. DOI: 10.48550/arXiv.2206.08219.

[18] V. Riego del Castillo, *Almost*, Jan. 2023. [Online]. Available: https://open.scayle.es/dataset/hand-landmarks-for-sos-gesture.

[19] I. Kotseruba and J. K. Tsotsos, "40 years of cognitive architectures: Core cognitive abilities and practical applications," *Artificial Intelligence Review*, vol. 53, no. 1, pp. 17–94, 2020.

[20] N. Mavridis, "A review of verbal and non-verbal human–robot interactive communication," *Robotics and Autonomous Systems*, vol. 63, pp. 22–35, 2015.

[21] M. Á. González-Santamarta, F. J. Rodrıéguez-Lera, C. Álvarez-Aparicio, Á. M. Guerrero-Higueras, and C. Fernández-Llamas, "Merlin a cognitive architecture for service robots," *Applied Sciences*, vol. 10, no. 17, p. 5989, 2020.

[22] C. Lugaresi, J. Tang, H. Nash, *et al.*, *MediaPipe: A Framework for Building Perception Pipelines*, arXiv:1906.08172 [cs], Jun. 2019. DOI: 10.48550/arXiv.1906.08172.

[23] F. Zhang, V. Bazarevsky, A. Vakunov, *et al.*, *MediaPipe Hands: On-device Real-time Hand Tracking*, arXiv:2006.10214 [cs], Jun. 2020. DOI: 10.48550/arXiv.2006.10214.

[24] G. Jocher, *YOLOv5 by Ultralytics*, version 7.0, May 2020. DOI: 10.5281/zenodo.3908559. [Online]. Available: https://github.com/ultralytics/yolov5.

[25] V. Riego del Castillo, *ROSE*, Jan. 2023. [Online]. Available: https://github.com/uleroboticsgroup/ROSE.

[26] Q. Gao, J. Liu, and Z. Ju, "Hand gesture recognition using multimodal data fusion and multiscale parallel convolutional neural network for human–robot interaction," *Expert Systems*, vol. 38, no. 5, e12490, 2021. DOI: 10.1111/exsy.12490.

[27] M. Peral, A. Sanfeliu, and A. Garrell, "Efficient Hand Gesture Recognition for Human-Robot Interaction," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 272–10 279, 2022. DOI: 10.1109/LRA.2022.3193251.