# A Framework for Real-Time Physical Human-Robot Interaction using Hand Gestures

Osama Mazhar, Sofiane Ramdani, Benjamin Navarro, Robin Passama, Andrea Cherubini

# A Framework for Real-Time Physical Human-Robot Interaction using Hand Gestures

Osama Mazhar Sofiane Ramdani Benjamin Navarro Robin Passama Andrea Cherubini

***Abstract*— A physical Human-Robot Interaction (pHRI) framework is proposed using vision and force sensors for a two-way object hand-over task. Kinect v2 is integrated with the state-of-the-art 2D skeleton extraction library namely *Openpose* to obtain a 3D skeleton of the human operator. A robust and rotation invariant (in the coronal plane) hand gesture recognition system is developed by exploiting a convolutional neural network. This network is trained such that the gestures can be recognized without the need to pre-process the RGB hand images at run time. This work establishes a firm basis for the robot control using hand-gestures. This will be extended for the development of intelligent human intention detection in pHRI scenarios to efficiently recognize a variety of static as well as dynamic gestures.**

## I. INTRODUCTION

For a successful and safe pHRI, appropriate understanding of the human user is essential for the robot. Increasingly popular and affordable consumer standard depth cameras like Microsoft Kinect has enabled the computer-vision and robotic researchers to develop robust pHRI systems in dynamic workplaces. It is well known that 93% of the human communication is non-verbal [1], of which 55% is accounted for elements like facial-expressions, posture, etc. In this perspective, ability to recognize human gestures can be extremely useful for a robotic system in pHRI scenarios [2].

Similar human-robot interaction (HRI) settings are found in the literature, such as [2], where the authors have used Kinect for the detection of pointed direction by the user and for navigation of the robot. In [3], the authors have used multiple Kinects in a fixed workspace and have used neural networks to detect dynamic gestures. Kinect based object recognition through 3D gestures is proposed in [4]. The OpenNI and NITE middleware are used to extract skeleton information of the human user. Authors in [5] and [6] also propose HRI scenarios using Kinect. Mostly the researchers have used OpenNI or Microsoft SDK to extract the human-skeleton, which is a model based skeleton tracker having several discrepancies including the need of initialization pose, not being able to detect gestures on which the model is not trained on, and noisy detections. Moreover, most of the gesture/intention detection work is done in the perspective of Human-Computer Interface (HCI) while pHRI implementations are not discussed.

In this paper, we integrate the state-of-the-art 2D skeleton extraction library namely *Openpose* [7] with Microsoft Kinect v2, which is a time-of-flight sensor, to get a real-time 3D skeleton of the human user. We also develop and train a convolutional neural network (CNN) for on-line hand gesture detection to control the robot. Moreover, a robot control framework is developed that combines vision and force sensors to achieve a two-way object handover between a human operator and a robot. The overall proposed pHRI scenario is shown in Fig. 1.

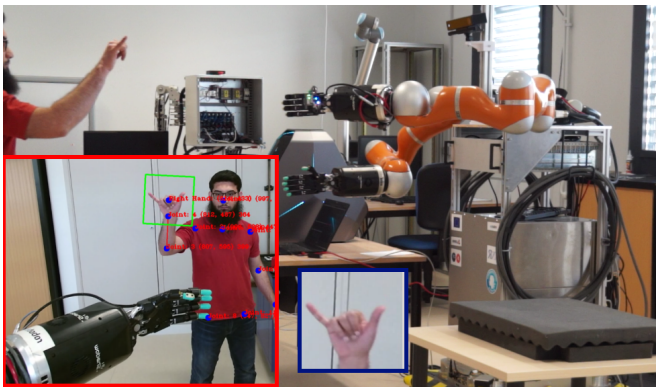

Fig. 1. Overall proposed pHRI scenario with BAZAR - The dual-arm mobile manipulator developed at LIRMM.

## II. METHODOLOGY

The proposed pHRI framework is divided in three main modules namely: *Skeleton extraction and hand image acquisition, CNN for pHRI hand gestures recognition* and *Robot control for pHRI*.

### A. Skeleton Extraction and Hand Images Acquisition

OpenPose is based on *Convolutional Pose Machines (CPMs)* [8] which extracts 2D skeleton information from RGB images. We use libfreenect2 library to get RGB image and depth map from Kinect v2, and extract the depth values of each 2D skeletal joint obtained from OpenPose, thus acquire a 3D skeleton of the human user. The angle of forearm is computed by fitting a line between the elbow joint and the wrist joint obtained using OpenPose. This line is then extended to one-third of its length to approximate the hand position. The angle that this line makes with horizontal, and the mean depth value of 36 pixels (6×6 matrix) of the wrist joint is used to fit a rotated bounding box centered at the approximated hand location. This enables our vision system to acquire a square image aligned with the forearm, of the size relative to the human hand, irrespective of the distance of the human operator from the camera within Kinect v2 depth sensing range. We crop, rotate and zoom the pixels lying within this bouding box to a size 244×244 pixels. In the detection phase, the hand cropped images are passed through

All authors are with LIRMM, Université de Montpellier, CNRS, Montpellier, France. osama.mazhar@lirmm.fr

the trained network without performing any time consuming image processing operation for gesture detection.

### B. CNN for Hand Gestures Recognition

We develop a CNN and train it on four gestures namely *Handover, Stop, Resume* and *None*. The architecture of our CNN is mainly inspired from LeNet [9] and is shown below:

INPUT→CONV(RELU:6×6×64)→MAXPOOL(2×2)
→DROPOUT(0.3)→CONV(RELU:3×3×128)
→MAXPOOL(2×2)→DROPOUT(0.5)→FC(128)
→FC(128)→FC(128)→SOFTMAX(4)

To make our CNN background invariant in an indoor environment, we used the depth map from the Kinect v2 to augment the data by removing the background from the training set of hand images. We also create an inverted binary mask of the hand to add different colors to the background in the training set. Keras data-augmentation is also used to introduce contrast stretching, channel shift and horizontal flip in the training data. The CNN is trained overnight on a set of 1800 RGB images of size 244×244 on an Intel Core i7-6800K CPU at 3.40GHz, 12 cores with no GPU. Validation accuracy, with 600 test images, is 98.8 %. We evaluated our model with 300 more test images extracted from a video recorded in different light conditions, and achieved 95.7 % accuracy on these data. We plan to extend this system by collecting more data from multiple users and to test the accuracy of our network on persons not included in the data.

### C. Robot Control for pHRI

The BAZAR robot used for the experiments is composed of two Kuka LWR 4+ arms with two Shadow Dexterous Hands attached at the end-effectors. The arms are attached to a Neobotix MPO700 omnidirectional mobile platform. In our scenario, the mobile base is kept fixed and only the right hand-arm system is used. The control of the arm is done using the FRI library and the control of the hand is based on a ROS interfaces. The external force applied to the arm's end-effector is estimated by FRI based on joint torque sensing and on knowledge of the robot's dynamic model. The control rate is set to 5ms.

## III. PHRI Experiment and Results

For safe pHRI, the robot must perceive the intention of the operator. Here, 3D human body joint coordinates and hand gesture recognition are the cues used for robot operation. We realize a tool (here, a portable screw-driver) handover experiment, guided by a finite state machine designed for the robot control. The robot waits for the user commands in the form of hand gestures, to take and then place the tool to a predefined location in its workspace. Once the handover gesture is detected, the robot moves to a predefined open hand position in the workspace. The human operator places the tool in the robotic hand and applies a downward force (in X-direction) on the end-effector to trigger tool grasping.

The experiment is demonstrated in a video that can be accessed through this link [1]. The commands are fulfilled by the robot when three successive identical instances of the corresponding gesture are detected, and only if the forearm is in the upper two quadrants of the axes centered at the elbow joint. This aids in ignoring all gesture detections when the operator does not intend to interact with the robot and has relaxed his/her arm.

This experiment is performed indoor and all gesture permutations are tested. The operator moves closer and farther from the robot and is allowed to move his hand in the coronal plane depending on his comfort. The robot is able to detect and obey the intended commands given by a single operator within approximately 384 milliseconds after the first instance of command is detected.

[1] www.youtube.com/watch?v=Mj5YqTDrdb4

## IV. Conclusion

Our current HRI framework detects hand gestures with a frame-rate of approximately 5.2 fps. The use of multiple GPUs for OpenPose library can enhance the temporal performance of our system. We explain our presented work in more detail in [10]. We plan to extend our work by developing a background independent hand gesture detector by substituting backgrounds with rich-textured images. This substitution makes the gesture detection a complex problem, thus we plan to exploit the concept of transfer learning in CNN to train our gesture detector. We also plan to integrate OpenPose asynchronously to the gesture detector to ensure faster execution of the algorithm.

## Acknowledgments

This work was supported by the CNRS PICS Project MedClub.

## References

[1] A. Mehrabian. *Nonverbal Communication*. Aldine Publishing Company, 1972.

[2] 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*, 149:65–77, 2016.

[3] G. Cicirelli, C. Attolico, C. Guaragnella, and T. DOrazio. A Kinect-Based Gesture Recognition Approach for a Natural Human Robot Interface. *Int. Journal of Advanced Robotic Systems*, 12(3):22, 2015.

[4] J. L. Raheja, M. Chandra, and A. Chaudhary. 3d gesture based real-time object selection and recognition. *Pattern Recognition Letters*, 2017.

[5] K. Ehlers and K. Brama. A human-robot interaction interface for mobile and stationary robots based on real-time 3d human body and hand-finger pose estimation. In *IEEE Int. Conf. on Emerging Technologies and Factory Automation (ETFA)*, pages 1–6, Sept 2016.

[6] Y. Yang, H. Yan, M. Dehghan, and M. H. Ang. Real-time human-robot interaction in complex environment using kinect v2 image recognition. In *IEEE Int. Conf. on Cybernetics and Intelligent Systems (CIS) and IEEE Conf. on Robotics, Automation and Mechatronics (RAM)*, pages 112–117, July 2015.

[7] Z. Cao, T. Simon, S. Wei, and Y. Sheikh. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017.

[8] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional Pose Machines. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2016.

[9] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998.

[10] O. Mazhar, S. Ramdani, B. Navarro, R. Passama, and A. Cherubini. Towards Real-time Physical Human-Robot Interaction using Skeleton Information and Hand Gestures. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, IROS*, 2018 (to appear).