ESAIM 2024 – 2nd European Symposium on Artificial Intelligence in Manufacturing

Human-Robot Interaction through Egocentric Hand Gesture Recognition

Snehal Walunj[1,*], Nazanin Mashhaditafreshi[2], Parsha Pahlevannejad[1], Achim Wagner[1], and Martin Ruskowski[2]

[1] German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany
[2] Rheinland-Pfälzische Technische Universität (RPTU) Kaiserslautern-Landau, Germany

[*] Corresponding author. Tel.: +49-631 20575-7078; E-mail: snehal.walunj@dfki.de

Abstract. Recognition of human hand gestures in industrial environments is gaining popularity, especially in the context of assistance-systems, thanks to advancements in AI-based vision methods. Also, head-worn devices with cameras are becoming more popular especially for smart assistance using Extended Reality (XR) technology, even for industrial use cases. Employing sensors from head-worn devices such as HoloLens enhance the communication between human and robot hereby providing interaction using ego-centric vision. This study delves into human-robot interaction by investigating ego-centered hand gesture recognition for commanding robots. A pipeline is developed for collecting these HoloLens video frames and to detect hand landmark labels on them using MediaPipe library by Google. Then, a Long Short-Term Memory Network (LSTM) model for hand-gesture recognition was developed that classifies the hand-gesture from the given hand landmarks in near real-time, which can then be translated into robot commands. We also present results for our network's performance and implementation pipeline using ROS communication.

Keywords: Egocentric Gesture Recognition · Hand Gesture Recognition · Human-Robot Interaction.

1 Introduction

Hand gestures are a natural and intuitive method humans use to communicate. These human-gestures can be translated into related robot commands [14] and enable hand gesture-based interface for human-robot interaction (HRI). Gesture-based control is a type HRI system that allows a human worker to control the robot's movements using gestures in a factory environment. Vision-based recognition systems enable workers to command robots which offers exciting possibilities for collaboration between human workers and machines [15]. Extended reality (XR) is also gaining popularity in the context of worker assistance system on account of its ability to augment information on to real-world to support human workers. These XR based assistance systems can be coupled with a robotic assistance system [6], [11], which together can supports workers in a Smart Factory environment. The head-mounted devices such as HoloLens2 consist of a camera that provide the egocentric view [13], [10] of the person using it. This data could potentially be used for Robot Interaction. Since it is a head mounted camera, it is capable of providing a constant data from the moving human, unlike a fixed camera. Also unlike Robot mounted cameras where the human must necessary be in the field-of-view, robots and humans can collaborate and co-operate at various scenarios in smart factories. However, the aim of this research is to leverage hand gesture to interact with robots in smart-factories using First-Person-View(FPV) cameras. Firstly, we need a method for data collection. Second, after collecting the required data, a machine learning model to classify the hand gestures should be developed. So, in the end, a pipeline can be established that includes data collection module, hand-gesture recognition module, robot-communication module.

2 Literature Review

Ambient cameras or fixed cameras (third-person view) offer the advantage of observing the entire scene, including the humans, their full-body gestures, and their environment. However, ambient camera-based hand gesture recognition (HGR) systems are often restricted by the sensor's range, requiring users to be near and/or directly in front of the cameras to perform gestures. On the contrary, wearable camera-based HGR systems overcome this limitation due to their portability [7], [2]. Wearable cameras such as head mounted cameras provide an egocentric view of the users that makes the users easily obersvable. However egocentric videos come with unpredictable movements and low quality due to constant motion blurs, which has been tackled using slow and fast pathways based on the frame rates of input videos in [2]. Other than only RGB sensors, different kinds of data sources such as depth information can be used to enhance the quality of recognition. In [5], authors proposed a novel architecture which combines RGB and Depth modalities evaluated on MECCANO dataset [12] that contains various hand gestures to mimic industrial settings. The paper by [8] introduces a mobile humanoid robot that can assist humans in public spaces by following hand gestures recognised using RGBD data from a robot mounted camera. Robot Operating System (ROS) which is an open source platform for robot control is used as middleware [8]. It consists of modules for human detection, tracking, and gesture recognition, based on neural network architectures. The paper [9] focuses on development and evaluation of a multistage spatial attention-based neural network for hand gesture recognition which are gaining popularity. There two type of gestures , the static ones and the dynamic ones [2], also they can be further classified as single-handed or two-handed gestures. Static gestures can be easily detected using Mediapipe library for hand pose tracking in real-time [16]. Mediapipe is used with a supporting algorithm based on Support Vector Machine (SVM)for hand-gesture recognition in [4]. It is possible to detect static gestures using a single image frame and libraries such as Mediapipe, however in dynamic gestures the hand poses vary with time, hence for which Recurrent Neural Networks (RNN) are used which considers the relationship between consecutive frames. LSTMs are type of RNNs that capture spectral, spatial as well as temporal features in a dataset [3]. Thus, it would be interesting to explore LSTMs together with CNN based models for static as well as dynamic hand gestures using single or both hands for real-time interaction with a robot, especially in a smart factory setting.

3 Implementation

The objective of this research is to develop and implement a comprehensive workflow for human-robot interaction using a combination of dataset creation, deep-learning model training, and evaluation on text data. A method for recognizing and classifying hand gestures in an egocentric perspective is being developed, utilizing a Long Short-Term Memory (LSTM) model, with the Microsoft HoloLens2 device camera. Our aim is to enable seamless communication between humans and robots, with a focus on recognizing and responding to human gestures in a factory setting. The project also emphasizes the importance of data pre-processing, evaluation on both recorded and real-time data. Alongside this work, an effort is made to implement gesture-based interaction between the factory worker and the robot using ROS communication.

3.1 Data Collection

Three classes of gestures needed to be classified. The "stop", "continue", and "come" hand gestures are defined, all being dynamic gestures. The "stop" and "come" gestures are both-handed, whereas the "continue" gesture could be done with either of the hands. The intention behind this setting was to have diverse type of gesture to be recognised. For which 1000 short videos of 39 frames for each class are recorded for the training dataset. For single handed gesture, around 500 samples were collected using the right hand, and the rest were collected using the left hand.

An important aspect of dataset collection process is to what extent can be automated,in order to save manual efforts. Since supervised-learning based deep learning models depend heavily on data. And large quantity of data needs to be collected in least possible efforts and time. For collecting egocentric data, Microsoft HoloLens2 is used. To save video data for further processing on a computer using HoloLens2 Sensor Streaming [1] application is utilized which transmits sensor data via TCP.

During data collection phase, different randomization techniques were included such as variable hand poses. Lighting conditions were also varied, as shown in Figure 1. The RGB data was not relevant in dataset collection phase, since the RGB data is used by the Mediapipe to give hand landmarks. Moreover, during certain data collection sessions, intentional hand tilts were introduced to simulate real-world scenarios and data with motion blur was also introduced, as depicted in Figure 1. Hand overlapping or occlusions is common, so various hand overlapping conditions were included. Moreover, the dataset was collected from 8 different individuals with varying hand sizes and gesture styles. These efforts aim to enhance future model training and improve the model's robustness and generalization.

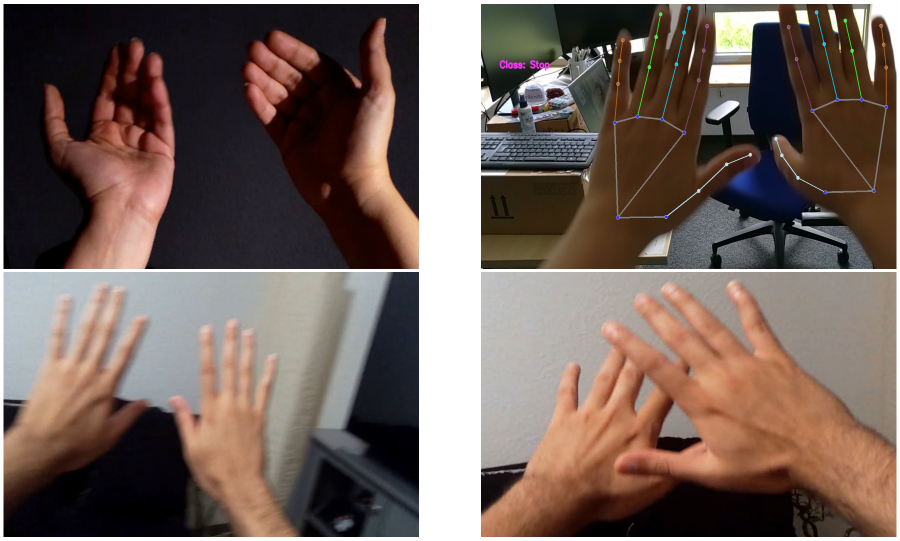

Fig. 1. Recorded data samples with different conditions of illumination, backgrounds, angles of hand poses and occlusion.

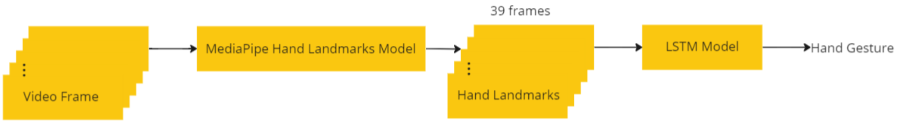

Fig. 2. Hand-gesture recognition pipeline

Throughout this work, two primary concerns need to be addressed. Firstly, the model must be lightweight to fulfill real-time requirements, allowing it to run efficiently on the robot. Secondly, the aim is to define our own set of gestures, necessitating the creation of a custom dataset and data collection method.

MediaPipe[3] already propose a CNN based approach trained on high quality and diverse dataset for hand landmark detection. It also performs very well and meet the real-time requirements. Based on the mentioned concerns and the fact that the MediaPipe hand-landmark detection model is powerful enough, a modular hand-gesture recognition approach, consisting of two modules, considered in this project. First, recorded video frames are fed to the MediaPipe CNN model to extract the hand keypoints, and save the keypoints for each frame. Then, these keypoints will be fed into a LSTM model for gesture detection. Each video contains 39 Numpy arrays of hand landmarks.Thus a complete gesture recognition pipeline was chaled out as shown in Figure 2.

3.2 Hand Landmarks

In the feature extraction phase, MediaPipe is used to extract hand landmarks. As illustrated in Figure 3, MediaPipe applies a robust model based on CNN to determine the keypoint localization of 21 hand-knuckle coordinates inside the detected hand regions. Mediapipe model was trained using around 30K real-world images as well as synthetically created hand models with a variety of backgrounds [16]. A palm detection model and a hand landmarks detection model are included in the MediaPipe hand landmarker model bundle. The palm detection model detects hands inside the input image, while the hand landmarks recognition model recognizes specific hand landmarks on the palm detection model's cropped hand image.

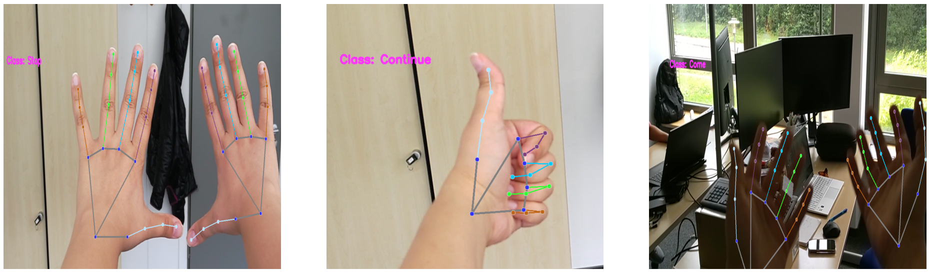

Fig. 3. From left to right, the stop, continue, and come hand gesture data labelled with hand landmarks

3.3 Classification Model

Recurrent Neural Networks (RNNs) are a specific type of neural network designed to retain information from previous states, and consider context and

[3] https://github.com/google/mediapipe

dependencies between time steps. The difference between RNNs, Long Short-Term Memory (LSTM) networks, and Gated Recurrent Unit (GRU) networks lies in the manner in which they manage memory and dependencies between time steps. To train an RNN, the network is back-propagated through time, and at each step, the gradient is calculated. If the gradient of the previous layer is smaller, the gradient of the current layer will be even smaller. This causes the gradients to shrink exponentially as they are back-propagated, which is known as the vanishing gradient, or short-term memory. To overcome this issue, a specialized version of the RNN was developed: Long Short-Term Memory (LSTM) network. LSTMs are highly effective at storing and accessing long-term dependencies, whereas RNNs are better suited to short-term dependencies and learn more quickly.

For the classification model a LSTM model was used. A stacked LSTM architecture was created, meaning that several LSTM units were concatenated. SGD optimizer with learning rate = 0.001 used as optimizer and the activation function is softmax for last dense layer and ReLu for other units. Also Categorical Cross-Entropy was used as the loss function. The corresponding model in block diagram is shown in Figure 4.

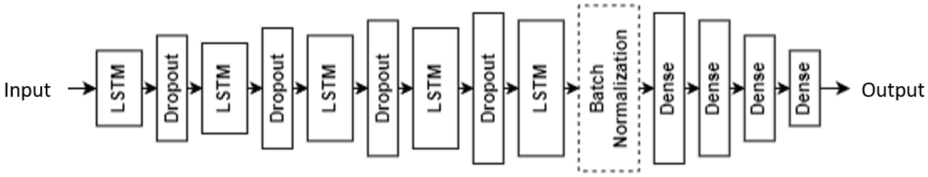

Fig. 4. Model architecture block diagram

4 Results

The final dataset was divided into 2186 samples for training, 243 for validation, and 608 for testing. The training phase is for 200 epochs, however after 10 epochs without significant improvement, training phase will stop. The model was trained in total for 82 epochs. The final test loss is 0.1287 and accuracy is 96%. In Figure 5, the model loss and accuracy is depicted over all epochs. The loss over all epochs is shown in Figure **??**. From the graphs, it can be derived that the training happens without overfitting or underfitting.

In table 1 the precision, recall, and f1-score is depicted for each class on the test dataset. The performance of all classes are mostly similar with f1-score being 0.97 for ‘stop‘, 0.98 for ‘continue‘, and 0.95 for ‘come‘.

The confusion matrix, as depicted in Figure 6, summarize the performance of the classification model. Based on the confusion matrix, the model performs

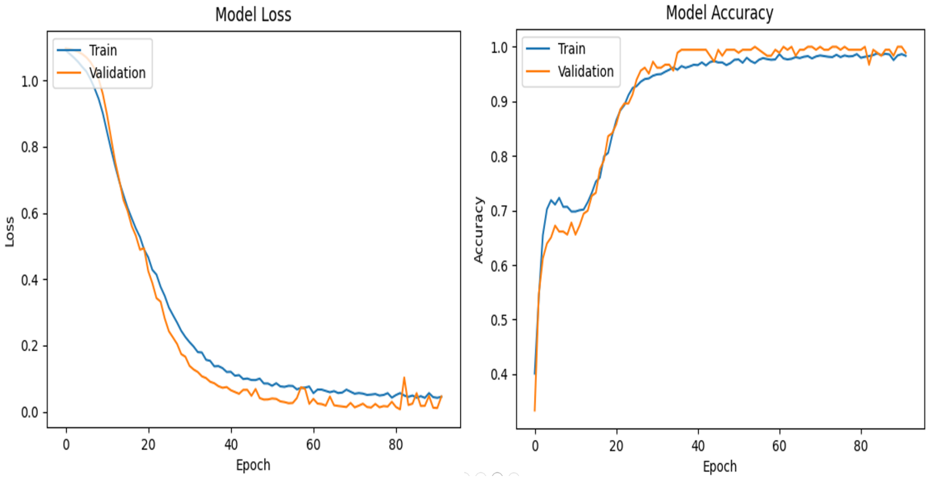

Fig. 5. Model loss and accuracy for train and validation set in 82 epochs.

Label	Precision	Recall	F1-Score	# Samples
Stop	1.00	0.93	0.96	439
Continue	0.96	1.00	0.98	390
Come	0.93	0.96	0.95	386

Table 1. Classification report on test dataset.

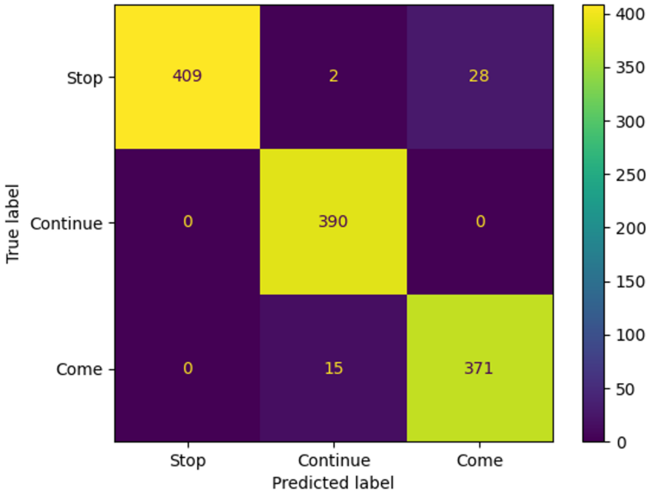

Fig. 6. Confusion matrix on test dataset.

better on ‘stop‘ and ‘continue‘ in comparison to ‘come‘. For the ’come’ gesture there was false classification as ’stop’ gesture. This can me improved with more data.

5 Robot Communication

In order to send commands to robot and inform the robot about worker’s requests, ROS communication was selected. 4 illustrates how the detected hand gestures are communicated to the robot. A publisher node has been developed on the PC, which publishes the recognized hand gestures into a ROS topic "HandGesture". On the other side, a subscriber node was also implemented, which will be informed in case of new messages in the topic. Depending on the received message the corresponding action will be executed on the robot.

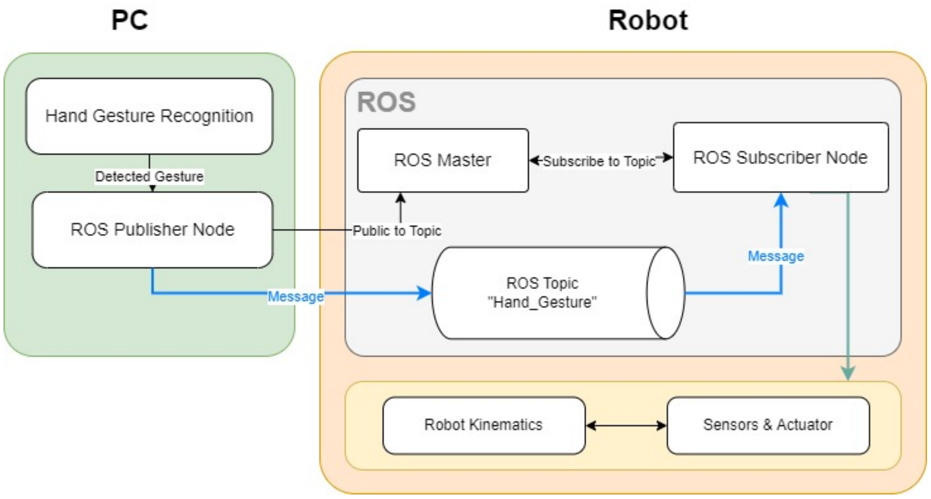

Fig. 7. ROS communication to send commands based on recognised gestures to robot

6 Conclusion and Future Scope

In this research we explored the potential of ego-centric hand gesture-based human robot interaction.Google’s MediaPipe library was used as the basis for extracting hand landmarks. These sequences of hand landmarks were then used for a classification of dynamic hand-gestures using an LSTM neural network. The performance of the model was promising (accuracy= 0.96, loss = 0.1287) and both the total accuracy as well as classification report, for each class as an evaluation metric, were considered. Although the results seem good, the amount of false positive classification of come gesture with stop gesture can be corrected using more dataset. Also experiments with videos with varying lengths could be used for training. More classes could be added to see how the model performs on more classes. Additionally, the potential of this gesture recognition model for hand-pose based action and activity recognition can be explored.

References

1. Dibene, J.C., Dunn, E.: Hololens 2 sensor streaming (2022)
2. Ho, H.D., Nguyen, H.Q., Nguyen, T.B., Vu, S.T., Le, T.L.: Dynamic hand gesture recognition from egocentric videos based on slowfast architecture. In: 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). pp. 01–07. IEEE (2022)
3. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
4. Kavana, K., Suma, N.: Recognization of hand gestures using mediapipe hands. International Research Journal of Modernization in Engineering Technology and Science **4**(06) (2022)
5. Kini, J., Fleischer, S., Dave, I., Shah, M.: Egocentric rgb+ depth action recognition in industry-like settings. arXiv preprint arXiv:2309.13962 (2023)
6. Kyaw, A.H., Spencer, L., Lok, L.: Human–machine collaboration using gesture recognition in mixed reality and robotic fabrication. Architectural Intelligence **3**(1), 11 (2024)
7. Le, V.D., Hoang, V.N., Nguyen, T.T., Le, V.H., Tran, T.H., Vu, H., Le, T.L.: A unified deep framework for hand pose estimation and dynamic hand action recognition from first-person rgb videos. In: 2021 International Conference on Multimedia Analysis and Pattern Recognition (MAPR). pp. 1–6. IEEE (2021)
8. Lindner, T., Wyrwał, D., Milecki, A.: An autonomous humanoid robot designed to assist a human with a gesture recognition system. Electronics **12**(12), 2652 (2023)
9. Miah, A.S.M., Hasan, M.A.M., Shin, J., Okuyama, Y., Tomioka, Y.: Multistage spatial attention-based neural network for hand gesture recognition. Computers **12**(1), 13 (2023)
10. Precup, S.A., Walunj, S., Gellert, A., Plociennik, C., Antony, J., Zamfirescu, C.B., Ruskowski, M.: Recognising worker intentions by assembly step prediction. In: 2023 IEEE 28th International Conference on Emerging Technologies and Factory Automation (ETFA). pp. 1–8. IEEE (2023)
11. Qian, L., Deguet, A., Wang, Z., Liu, Y.H., Kazanzides, P.: Augmented reality assisted instrument insertion and tool manipulation for the first assistant in robotic surgery. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 5173–5179 (2019). https://doi.org/10.1109/ICRA.2019.8794263
12. Ragusa, F., Furnari, A., Farinella, G.M.: Meccano: A multimodal egocentric dataset for humans behavior understanding in the industrial-like domain. Computer Vision and Image Understanding (CVIU) (2023). https://doi.org/https://doi.org/10.48550/arXiv.2209.08691, https://iplab.dmi.unict.it/MECCANO/
13. Tavakoli, H., Walunj, S., Pahlevannejad, P., Plociennik, C., Ruskowski, M.: Small object detection for near real-time egocentric perception in a manual assembly scenario. arXiv preprint arXiv:2106.06403 (2021)

14. Villani, V., Secchi, C., Lippi, M., Sabattini, L.: A general pipeline for online gesture recognition in human–robot interaction. IEEE Transactions on Human-Machine Systems **53**(2), 315–324 (2023)
15. Walunj, S., Sintek, M., Pahlevannejad, P., Plociennik, C., Ruskowski, M.: Full paper: Ontology-based digital twin framework for smart factories (2023)
16. Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.L., Grundmann, M.: Mediapipe hands: On-device real-time hand tracking (2020)