ICCEPM 2022
The 9th International Conference on Construction Engineering and Project Management
Jun. 20-23, 2022, Las Vegas, NV, USA

Multi-camera-based 3D Human Pose Estimation for Close-Proximity Human-robot Collaboration in Construction

Sajib Sarkar[1]*, Youjin Jang[2], Inbae Jeong[3]

[1] *Department of Civil, Construction and Environmental Engineering, North Dakota State University, 1410 North 14th Avenue, Fargo, ND 58102, USA,* E-mail address: sajib.sarkar@ndsu.edu
[2] *Department of Civil, Construction and Environmental Engineering, North Dakota State University, 1410 North 14th Avenue, Fargo, ND 58102,* E-mail address: y.jang@ndsu.edu
[3] *Department of Mechanical Engineering, North Dakota State University, 111 Dolve Hall, Fargo, ND 58108, USA,* E-mail address: inbae.jeong@ndsu.edu

Abstract: With the advance of robot capabilities and functionalities, construction robots assisting construction workers have been increasingly deployed on construction sites to improve safety, efficiency and productivity. For close-proximity human-robot collaboration in construction sites, robots need to be aware of the context, especially construction worker's behavior, in real-time to avoid collision with workers. To recognize human behavior, most previous studies obtained 3D human poses using a single camera or an RGB-depth (RGB-D) camera. However, single-camera detection has limitations such as occlusions, detection failure, and sensor malfunction, and an RGB-D camera may suffer from interference from lighting conditions and surface material. To address these issues, this study proposes a novel method of 3D human pose estimation by extracting 2D location of each joint from multiple images captured at the same time from different viewpoints, fusing each joint's 2D locations, and estimating the 3D joint location. For higher accuracy, the probabilistic representation is used to extract the 2D location of the joints, considering each joint location extracted from images as a noisy partial observation. Then, this study estimates the 3D human pose by fusing the probabilistic 2D joint locations to maximize the likelihood. The proposed method was evaluated in both simulation and laboratory settings, and the results demonstrated the accuracy of estimation and the feasibility in practice. This study contributes to ensuring human safety in close-proximity human-robot collaboration by providing a novel method of 3D human pose estimation.

Keywords: construction worker, human-robot collaboration, human pose estimation, close-proximity

1. INTRODUCTION

The construction industry plays a vital role in both developed and developing economies in terms of creating new jobs, driving economic growth, and providing solutions to address social, climate, and energy challenges. Nevertheless, the overall productivity of the construction industry is significantly lower compared to other key industrial sectors. The productivity growth in construction has averaged one percent a year over the past two decades, compared with growth of 2.8 percent for the total world economy and 3.6 percent in manufacturing [1]. As a result, the productivity in the construction industry lags far behind that of manufacturing or the entire economy. In addition to the stagnant productivity, the construction industry also suffers from a labor shortage [2]. Construction tasks are labor-intensive and physically demanding. These characteristics make it difficult to steer a new workforce to the construction industry. Also, a relatively high number of deaths and injuries at construction sites have been placed due to the unstructured, fast-changing environment of the construction site. In 2017, 971 fatalities were recorded at construction sites [3].

Automation and robotics in construction can improve productivity and address the labor shortage and safety problems. A robot can conduct certain construction tasks that are inherently repetitive or dangerous at a higher speed, power, and precision than a human counterpart. Also, a robot can deliver high volumes of construction materials in less time. However, there are still tasks that only human workers can do with the cognitive skills of humans. For example, when installing drywall, a robot has difficulty making decisions adaptively in unknown situations such as tuning the drywall panel or adjusting nailing angles while a human worker can quickly improvise a new plan in such a situation. Therefore, there is a need to combine the strength of human workers and robots. Close-proximity collaboration between human workers and robots is expected to create new synergies and tasks optimization opportunities and will lead to a new paradigm in the construction industry.

In close-proximity human-robot collaboration, ensuring the safety of human workers is a significant challenge. Collisions between human workers and robots may cause interruption of work, malfunction of robots, and severe injuries. Therefore, it is critical to accurately know where human workers and robots are in the workspace and subsequent actions they are going to take in real-time. Workers' location and their behaviors can be predicted from 3D human pose estimation, involving the estimated 3D locations of the human body. There have been studies to estimate human poses [4-10]. Most previous studies obtained 3D human poses using a single camera or an RGB-depth (RGB-D) camera. However, single-camera detection has limitations such as occlusions, detection failure, and sensor malfunction, and an RGB-D camera may suffer from interference from lighting conditions and surface material [11,12]. To address these problems, this study proposes a novel method of multi-camera-based 3D human pose estimation. Then, the simulation and laboratory experiments are conducted to demonstrate the feasibility, efficiency, and accuracy of the proposed model.

The remainder of this paper is organized as follows. First, the literature on human pose estimation is reviewed. Next, the proposed multi-camera-based 3D human pose estimation method is described, and the results of the simulation and laboratory experiments are demonstrated by the proposed methods. Finally, discussions and conclusions of the study are presented.

2. LITERATURE REVIEW

The skeleton model is widely utilized in human pose estimation in two dimensions and it is naturally extended to three dimensions. The human skeleton model contains numerous major locations of the human body and uses edges between key joints to connect natural nearby joints. Most prior studies collected 3D joint locations utilizing a single camera or an RGB-D camera to recognize human behavior. By matching the data retrieved by the camera with a predetermined human database, the whole stance of the human skeleton may be determined. However, in a single-camera detection, the position of particular human joints cannot be accurately deduced when the human subject is not fully visible to the camera. To overcome occlusion problem, multiple RGB-D cameras are employed to compute the distance from dynamic objects in real-time [4]. Multiple RGB-D cameras are simultaneously leveraged to minimize occlusion by combining the entire data obtained [5]. This practice, however, results in a massive volume of redundant data that is difficult to handle. Furthermore, the method necessitates a complex design to combine the many devices.

Multi-camera view-based investigations are challenging because multiple 3D predictions may produce the same 2D projection, making 3D pose estimation from monocular inputs an unsolved task. Researchers have used multi-view geometry for feature fusion and triangulation, or 2D convolutional neural networks for a human subject's joint detection and pictorial structural models for rapid and reliable 3D position reconstruction [6]. These approaches are computationally very expensive and because of the large volume grid, they are not real-time capable. Ragaglia et al. (2018) used a kalman filter, which estimate the internal-state of a linear dynamic system from a series of noisy measurements, to estimate human posture [7]. However, despite being computationally efficient and simple to construct approach, it was inappropriate for dealing with occlusions.

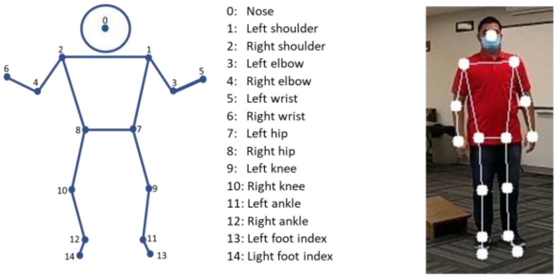
The particle filter is an alternate strategy for addressing human pose estimation that performs well in the presence of nonlinearities and non-Gaussian distributions. The particle filter is used to improve the accuracy of the RGB-D camera in the case of occlusions by combining color and depth data [8]. Casalino et al. (2018) also presented an algorithm for human pose estimation in the situations of partial occlusion based on particle filter techniques [9]. By considering anatomic distances of the human body and information received by an RGB-D camera, the suggested algorithm was effective in decreasing the uncertainty associated with the occluded human position.

However, an RGB-D camera may suffer from interference from lighting conditions and surface materials. RGB-D camera has a limited range of measurement and is prone to reflections on clear, glossy, or highly matte and absorbing materials. The infrared ray patterns interfere with each other if more than one RGB-D camera is used, therefore, a significant amount of depth information is lost. As such, multiple RGB-D cameras have limitations in using them in outdoor environments such as construction work. Therefore, this study uses multiple monocular cameras to estimate 3D human pose with a particle filter.

3. MULTI-CAMERA-BASED 3D HUMAN POSE ESTIMATION

3.1. Extracting 2D joint locations of a human

In order to extract the 2D joint locations on the image captured from a camera, an existing human pose estimation algorithm is used. As the performance of the 3D pose estimation process relies on the base 2D joint detection algorithm, the algorithm should have high accuracy and real-time performance. In this study, MediaPipe Pose was adopted, which is one of the state-of-the-art technologies with the capability of high-fidelity pose detection and real-time performance. MediaPipe Pose utilizes a two-step detector-tracker process that first locates the pose region-of-interest (ROI) in an image and then predicts the locations of joints on the image coordinate [10].



(a) Keypoints (b) Extracted 2D joint location

Figure 1. 2D joint locations of a human

A total of 15 keypoints are extracted out of 33 keypoints of the BlazePose detection model from the result of the underlying pose detection algorithm. The keypoints are used as the joint locations in the following 3D human pose estimation process. Figure 1 illustrates the 15 keypoints used in this study.

3.2. Generating 3D human pose location using particle filter

This study utilizes multiple camera sensors that are connected through the network. Multiple camera sensors are located at known locations and orientations, and each camera sensor captures an image of the user in the environment. 2D joint locations of 15 keypoints, collected through MediaPipe Pose, are converted into 3D locations through rotation, translation, and homogeneous transformation. The underlying human pose detection algorithm is run to detect the human pose on each image captured at the same time. The proposed algorithm then gathers the pose estimated on each 2D image and fuses the information for accurate 3D human pose estimation. As the 2D pose estimation algorithm often fails to detect the human pose or generates the wrong pose estimate due to the occlusions by objects or other bodies, the 2D estimation results are considered noisy observation and are used as inputs of the probabilistic estimation algorithm.

```
Algorithm 1 EstimatePose(N, M, C)
Parameters
  N: number of iterations
  M: number of particles
  C: Set of location and orientation of camera sensors
1:  P = initialize_particle(M) // A particle has 3D locations ($x_i$,$y_i$,$z_i$) of i-th joint for i=0,…,14
2:  for n=1 to N do
3:    for i = 1 to n(C) do
4:      image[i] = capture_frame(C[i])
5:    endfor
6:    for m = 1 to M do
7:      P[m] += N(0, Σ) // Pose change as a free rolling ball
8:      W[m] = 0
9:      for i = 1 to n(C) do
10:       // Expectation and observation of 2D locations of joints on the image coord.
11:       $expected_{m,i}$ = project_2d(P[m], C[i])
12:       $observation_i$ = detect_pose_2d(image)
13:       W[m] = update(W[m], $expected_{m,i}$, $observation_i$)
14:     endfor
15:   endfor
16:   W = normalize(W)
17:   P = resample(P, W)
18: endfor
19: Return estimated pose
```

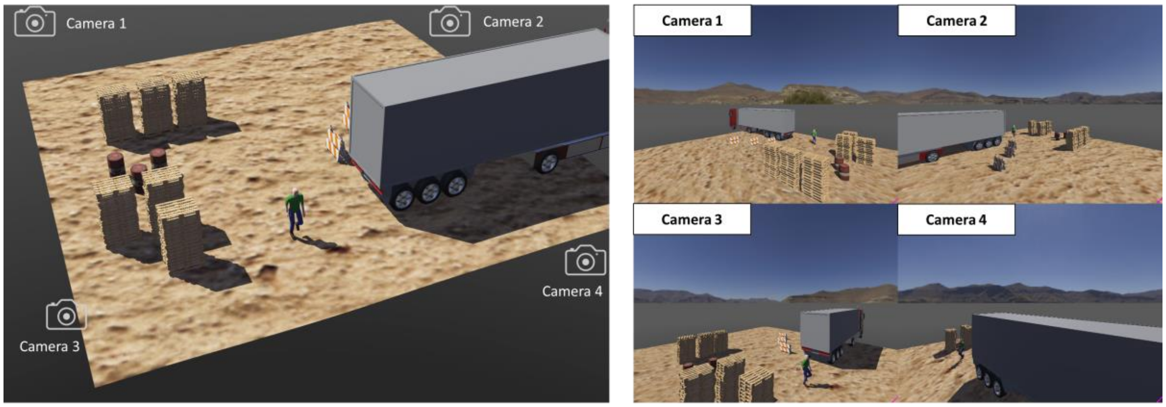
Figure 2. The proposed algorithm

The particle filter algorithm is adopted to fuse the observations from the camera sensors and estimate the 3D human pose. The particle filter algorithm is an algorithm for estimating unknown internal states from partial noisy observations. Figure 2 shows the particle filter algorithm implemented for the estimation process. It starts from a set of randomly distributed particles (Line 1). As human movement is unknown, the pose is modeled as a free rolling ball for each joint, which moves in a random direction (Line 7). The human pose in each particle is then projected to an image plane (Line 11) as if it is captured with i-th camera sensor, and the projected joint locations are compared to the observation (Line 12). For each pair of expected and observed joint locations on the image coordinate, the Euclidean distance between two locations are considered Gaussian random variable and are used to update the weight of each particle (Line 13). The weights of the particles are then normalized (Line 16), and the particles are resampled by the weights (Line 17). As a particle with high weight has a higher probability to be resampled, the particles are iteratively updated to present the posterior distribution of the 3D human pose.

4. EXPERIMENTAL RESULTS

4.1. Simulation experimental settings and results

To test the feasibility of the proposed 3D human pose estimation model, this study carried out simulation experiments using a robotics simulator. While the real-world experiment in the following section detects all joints, this simulation experiment detects only head, focusing on the comparison of the effectiveness of the multi-camera and single-camera approaches, to prove that the multi-camera approach does not suffer from detecting the human pose due to occlusion. Webots is an open-source 3D robot simulator that has a range of actuators, sensors, and objects which realistically mimic hardware components in the real world and supports their physical properties such as gravity, friction, and dynamics. In this study, a rectangular-shaped virtual space of 20m × 20m × 10m, consisting of construction objects such as a truck, wood pieces, etc., was used as a simulation environment. Four cameras were located at each corner at a height of 5m and captured images of human subjects who walked in the simulation environment (see Figure 3).

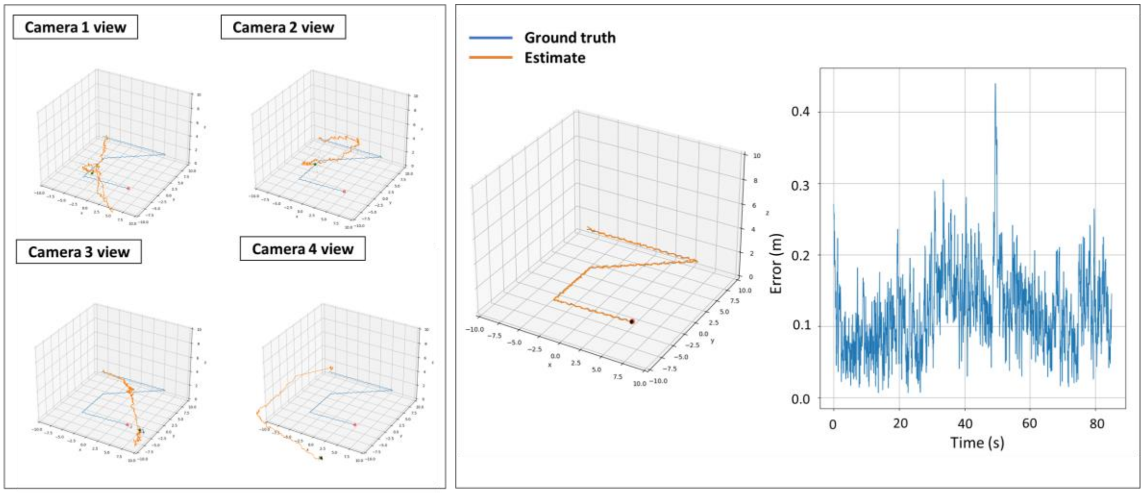


(a) Location of Cameras and Human Subject

(b) Simulation Experiment Environment

Figure 3. Simulation Experimental Settings

Figure 4 shows the trajectories of the ground truth and estimated location of the human subject's head, using a single camera and multiple cameras. When using cameras individually, the human poses were not estimated correctly. The root mean square errors (RMSE) of the estimation from the cameras 1,2,3, and 4 were 6.771, 3.763, 5.630, and 12.814, respectively. On the other hand, when multiple cameras were used with the proposed model, the RMSE of the estimation was 0.133. Therefore, the results of the simulation experiment demonstrate that the estimation accuracy of the proposed method in this study outperforms the single camera approach.



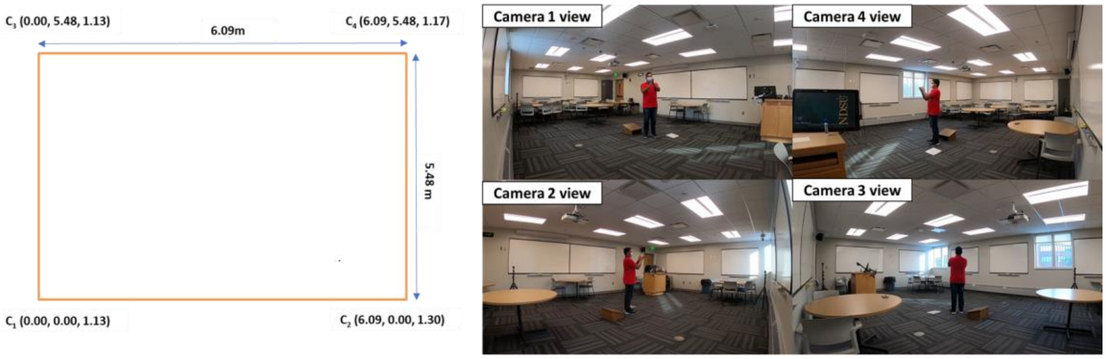
(a) 3D human pose estimation using a single camera

(b) 3D human pose estimation using multiple cameras

Figure 4. Results of Simulation Experiments

4.2. Laboratory experimental settings and results

In this experiment, in addition to the head joint as in the previous simulation environment, all joints were detected and their 3D locations were estimated using the same algorithm in parallel, to estimate the human subject's pose. This study conducted laboratory experiments in a rectangular-shaped space of 6.09m × 5.48m× 2.59m with a human subject. Four GoPro HERO8 Black cameras were set up at different locations to capture 1080p HD videos of s human subject's motions from four different angles, as shown in Figure 5. The wide lens was set as the camera's digital lens mode in which the horizontal field of view (FOV) is 118.2 degrees and the vertical FOV is 69.5 degrees. The locations of cameras (C_1, C_2, C_3, and C_4) for the world coordinate frame are shown in Figure 5(a).



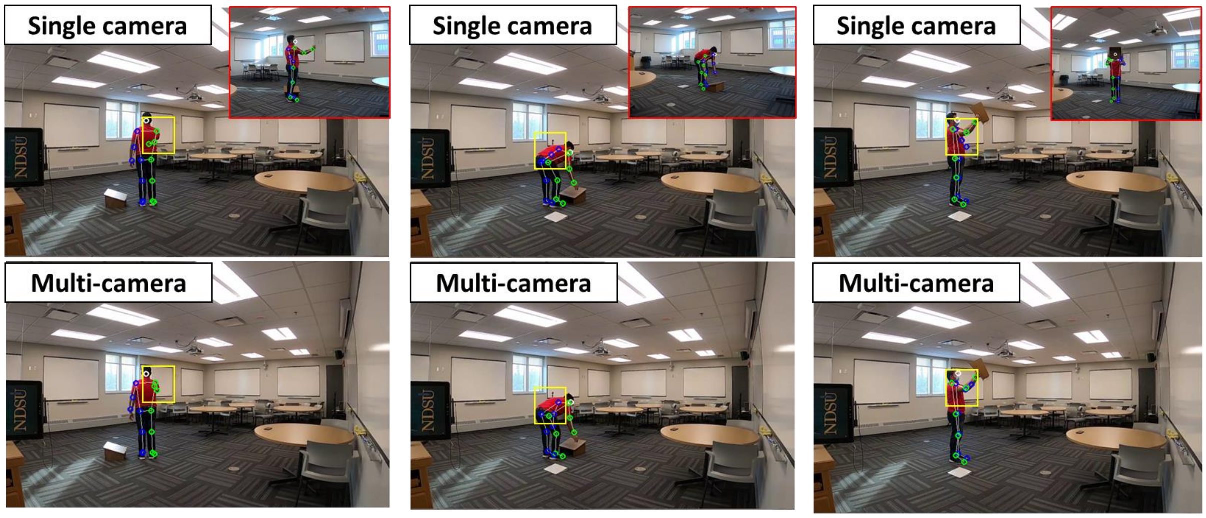
(a) Location of Cameras

(b) Laboratory Experiment Environment

Figure 5. Laboratory Experimental Settings

For the experiment, a human subject performed a variety of motions, including standing, walking, bending, lifting, and moving objects such as a whiteboard and wood piece. The motion of a human subject was captured from four different viewpoints. The captured images were grouped by the timestamp so that the images taken at the same time could be easily used in the joint estimation process.

Figure 6 compares the results of 3D pose estimation between using a single camera and multiple cameras. While a human subject is just standing, some right-sided body keypoints were not correctly detected and estimated when only a single camera was used, whereas it was possible to estimate the missing keypoints from the multi-views (see figure 6(a)). When the human subject was bending to grab a wood piece, a single camera view failed to estimate the joint locations of the left-sided keypoints of the human body while the proposed method could estimate them correctly (see figure 6(b)). In addition, left-side keypoints of the human body were not detected from a single camera when a human subject was lifting a whiteboard, but the proposed method could detect all joint locations by fusing the information extracted from the rest of the cameras (see figure 6(c)). The results of the laboratory experiment show the feasibility of the proposed method in practice.



(a) Standing (b) Bending (c) Lifting object

Figure 6. 3D pose estimation using a single camera and multiple cameras

5. CONCLUSIONS

This study provides a novel approach for 3D human pose estimation by fusing 2D joint locations extracted from multiple images from multiple viewpoints. A total of 15 key points were extracted and used as 2D joint locations of a human subject. The 3D human pose was then approximated using the particle filtering technique to fuse 2D joint locations probabilistically. Both simulation and laboratory experiments showed that a single camera works well to get the accurate 3D pose estimate only when the human subject was facing toward a camera. When the human subject makes a turn towards left or right or partially occluded, a single camera system often does not work well. This study shows the feasibility of the proposed method in practice. The proposed method is expected to be useful not only for human-robot collaboration in construction but also for any work environment where robots would be employed to work along with humans in close proximity. The robots should plan trajectories of their manipulators or navigation plans so that they never make any collisions with nearby humans. More accurate pose and location information of nearby humans provided by the proposed algorithm will enable identification of their behaviors and prediction of next motions, and therefore the robot will be able to use the information and calculate a safe plan that avoids making collisions with nearby humans. However, there is still room for improvement for adoption in the real world. This study conducted simulation and laboratory experiments in an enclosed space environment with only one person instead of a real construction field with multiple persons. Also, a limited number of poses of a human subject were tracked whereas there could be a wide variety of poses of construction workers. Future works include tracking the 3D human pose of multiple construction workers with various poses in the real construction site environment.

Furthermore, this study depends on the joint kinematics of the underlying single-camera detection algorithm. The performance of the proposed pose estimation approach can be further improved by taking joint kinematics into consideration. Lastly, the high-precision 3D joint location estimation achieved from this study can be used to develop an enhanced prediction model of human motions for human-robot collaboration in construction tasks.

ACKNOWLEDGMENTS

The authors would like to acknowledge financial support from the ND EPSCoR STEM grants program for this research.

REFERENCES

[1] F. Barbosa, J. Woetzel, J. Mischke, M. Ribeirinho, M. Sridhar, M. Parsons, N. Bertram, and S. Brown, "Reinventing Construction: A route to higher productivity." McKinsey Global Institute, 2017.

[2] Bureau of Labor Statistics, Job openings levels and rates by industry and region, seasonally adjusted. https://www.bls.gov/charts/job-openings-and-labor-tu

[3] Bureau of Labor Statistics, National Census of Fatal Occupational Injuries in 2017. https://www.bls.gov/news.release/pdf/cfoi.pdf

[4] F. Flacco and A. De Luca, "Real-time computation of distance to dynamic obstacles with multiple depth sensors", In: IEEE Robotics and Automation Letters, vol. 2, no. 1, pp. 56–63, 2016.

[5] N. Chen, Y. Chang, H. Liu, L. Huang, and H. Zhang, "Human pose recognition based on skeleton fusion from multiple kinects," In: 2018 37th Chinese Control Conference (CCC), IEEE, pp. 5228– 5232, 2018.

[6] R. Hartley and A. Zisserman, "Multiple View Geometry in Computer Vision",

In: Cambridge University Press, New York, NY, USA, 2 edition, 2003.

[7] M. Ragaglia, A. M. Zanchettin, and P. Rocco, "Trajectory generation algorithm for safe human-robot collaboration based on multiple depth sensor measurements", Mechatronics, vol. 55, pp. 267–281, 2018.

[8] A. Yoshida, H. Kim, J. K. Tan, and S. Ishikawa, "Person tracking on Kinect images using particle filter", Proceedings of the 2014 Joint 7th SCIS and 15th ISIS. IEEE, pp. 1486–1489, 2014.

[9] A. Casalino, S. Guzman, A. M. Zanchettin, and P. Rocco, "Human pose estimation in presence of occlusion using depth camera sensors, in human-robot coexistence scenarios", Proceedings of the 2018 IEEE/RSJ International Conference on IROS, IEEE, pp. 6117–6123, 2018.

[10] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.Chang, M. Yong, J. Lee, W. Chang, W. Hua, M. Georg, and M. Grundmann, "MediaPipe: A framework for building perception pipelines", arXiv preprint arXiv:1906, 08172, 2019

[11] Shih-Wei Sun, Chien-Hao Kuo, and Pao-Chi Chang. 2016. "People tracking in an environment with multiple depth cameras." J. Vis. Comun. Image Represent. 35, C (February 2016), 36–54.

[12] Alhwarin F., Ferrein A., Scholl I. (2014) IR Stereo Kinect: Improving Depth Images by Combining Structured Light with IR Stereo. In: Pham DN., Park SB. (eds) PRICAI 2014: Trends in Artificial Intelligence. PRICAI 2014. Lecture Notes in Computer Science, vol 8862. Springer, Cham.