# Evaluation of 3D markerless pose estimation accuracy using OpenPose and depth information from a single RGB-D camera

Fotios Lygerakis
University of Texas at Arlington
fotios.lygerakis@mavs.uta.edu

Athanasios C. Tsitos
NCSR Demokritos
tsitos@iit.demokritos.gr

Maria Dagioglou
NCSR Demokritos
mdagiogl@iit.demokritos.gr

Fillia Makedon
University of Texas at Arlington
makedon@uta.edu

Vangelis Karkaletsis
NCSR Demokritos
vangelis@iit.demokritos.gr

## ABSTRACT

Safe and efficient Human-Robot Collaboration (HRC) requires recognizing human collaborators intention. Hand pose kinematics early on an ongoing movement can provide information for predicting future human actions. Accurate state-of-the-art methods used for human hand pose estimation are either marker-based or make use of multiple cameras set around the workspace. These approaches introduce inconvenience to the user, necessitate calibration and are bounded to the specific set-up and workspace. On the other hand, using a single RGB-D camera would be less obtrusive for the user and less cumbersome to install. In this work, we use OpenPose to extract 2D keypoints from the RGB raw image and we combine them with the depth information acquired from the RGB-D camera to obtain 3D hand poses. We evaluate the accuracy and discrimination ability of our method in ten different static poses.

## CCS CONCEPTS

• **Human-centered computing → Gestural input**; *Usability testing*.

## KEYWORDS

human robot collaboration, 3D hand pose detection, RGB-D camera, OpenPose

**ACM Reference Format:**
Fotios Lygerakis, Athanasios C. Tsitos, Maria Dagioglou, Fillia Makedon, and Vangelis Karkaletsis. 2020. Evaluation of 3D markerless pose estimation accuracy using OpenPose and depth information from a single RGB-D camera. In *The 13th PErvasive Technologies Related to Assistive Environments Conference (PETRA '20), June 30-July 3, 2020, Corfu, Greece.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3389189.3398005

## 1 INTRODUCTION

The ever-growing integration of robots in the fields of home care, nursing, space exploration and rescuing missions, as well as their widespread use in every industrial environment has lead to the great need of studying the interaction between them and their users. Human Robot Collaboration (HRC) is the process during which a person and a robotic system, e.g. a robotic arm, are interacting physically, socially or both[6, 8, 23] in order to achieve a goal. Human environments, however, are complex and non-deterministic and, thus, user's safety during interaction is not a trivial challenge[10, 13, 21]. On the contrary, HRC involves constant position tracking, intention estimation and action prediction of the user.

User's motion intention estimation is a very important variable for the safe and efficient physical collaboration with a robotic system. A Collaborative Robot (Cobot) having the ability to assess user's intention [5], e.g. grasp a specific object[1], could take into consideration potential hazards and motion efficiency when planning motions. Consequently, it is important to obtain accurate and repeatable estimation of the user's hand pose [7].

We propose a novel use of a single RGB-D (Red Green Blue Depth) Camera to detect human 3D hand pose. The configuration we propose is totally adjustable to all collaborative robotic arms, since such set-ups use at least one camera to observe the workspace. The RGB-D camera can either be mounted on the robot or observe the workspace standalone. Our proposed method can also be used on every wheeled or legged mobile robot, since the camera can be mounted on the robot itself. For this reason we use a RGB-D camera that incorporates an inertial measurement unit (IMU), that can be efficiently integrated on a mobile Cobot to provide a wide spectrum of the needed functionalities.

## 2 RELATED WORK

In behavioral neuroscience and psychology studies, a popular and accurate method to estimate a user's 3D hand poses is based on the use of markers on the hand. In [1, 5] the hand poses are being detected with the use of reflective markers on the hand. However, marker-based methods[11] require precise adjustment of the markers on the hands of each individual. Additionally, wearing markers could be quite obtrusive for users in HRC set-ups.

Many markerless motion capture approaches, that could be employed to detect 3D hand poses, have been proposed so far[3, 14, 15, 17], providing state-of-the-art accuracy. Despite their independence of the use of markers, they still require the use of multiple cameras to extract 3D poses. Likewise, this scenario is not practical and entirely infeasible in case where the collaboration involves user's interaction with a mobile robot.

Markerless 3D hand pose estimation using a single RGB camera has also been proposed[18, 24]. Nevertheless, these single-camera


*PETRA '20, June 30-July 3, 2020, Corfu, Greece*

ACM ISBN 978-1-4503-7773-7/20/06…$15.00
https://doi.org/10.1145/3389189.3398005

RGB-based approaches suffer from performance degradation in comparison to the aforementioned state-of-the-art methods. In addition, the proposed model's performance in [24] is bounded by the lack of a large-scale appropriate dataset. Other neural-network based approaches include mapping from 2D RGB images to 2.5D hand pose representations to reconstruct the 3D hand poses[12] and kinematic model fitting with synthetic data enrichment such that it resembles the distribution of real hand image [16].

## 3 CONTRIBUTION

We integrate and evaluate the accuracy of a 3D hand pose extraction method based on OpenPose 2D hand pose estimation method [20]. Our method utilizes the accuracy and robustness of OpenPose model, while introducing simplicity at the 3D markerless hand pose estimation. A similar approach has been used for body pose tracking in [2] and both body and hand pose tracking in [2]. However, to our knowledge, there has not been any systematic evaluation of hand position estimation using similar methods. Yet, knowing the accuracy of such tools is crucial for designing safe and natural HRC environments that demand precise human motion tracking and human action prediction. The impact of achieving accurate human behaviour monitoring through a single, low-cost, commercial, RGB-D camera would be great given the wide applicability of the method, the flexibility in the set-ups, and the accessibility of the hardware.

## 4 METHOD

### 4.1 Participants, Apparatus and Workspace

In this study, a right-handed male participant was asked to sit on a height-adjusted seat in front of a table (length= 150cm, width= 70cm, height= 60cm) as shown in Figure 1. On the table an omnigrid mat with a square size of 2.54cm was placed, in order to assess the detected fingertip positions in comparison to the ground-truth ones.

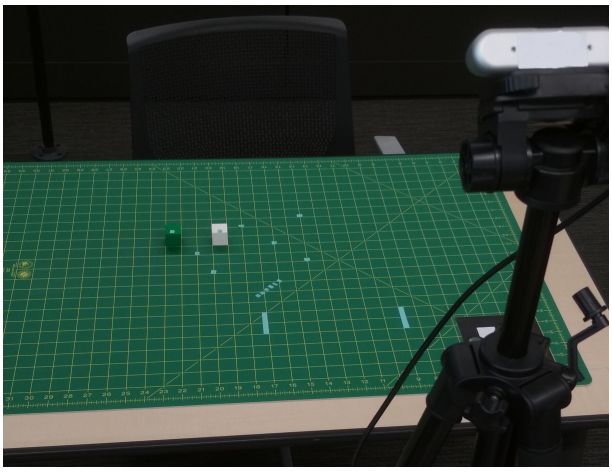


**Figure 1: Workspace set-up.**

In a distance of 50cm from the table we place a 150cm tall tripod on which we placed an Intel RealSense Depth Camera D435i (resolution: 1920 x 1080 at 30 fps). The pose of the camera with respect to the table was determined by identifying the pose of an individual augmented reality (AR) [1] marker. The center of the marker was placed at x=3 cm y=3 cm with respect to the table reference frame, which was defined as one of the corners of the omnigrid mat. These transformations were used in Frame Transpose node[2] (Figure 3) to transform all collected camera points from camera frame to table reference frame.

The chair's height had to be adjusted, so that the participant's neck and shoulders are also visible. Furthermore, sixteen stickers were added on the omnigrid mat in order to compare the proposed system's measurements with the ground-truth positions, as well as two cubes of edge size 2.54cm to provide different positions on the z-axis(height), as shown in Figure 2. Each big square of the grid is 2.54cm and the double squares represent the used cubes.

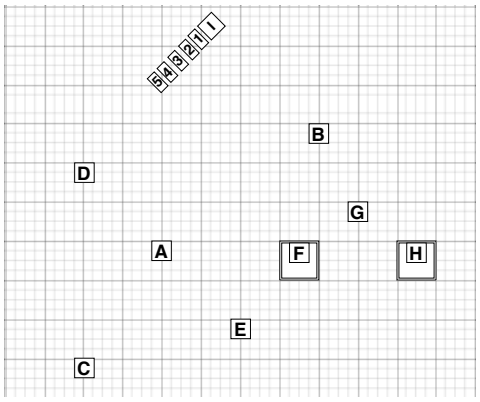



**Figure 2: The 16 stickers on the omnigrid mat. Double squares represents cubes.**

### 4.2 Data Collection

We defined ten static poses, for which we asked the participant to place his right-hand index and thumb on the specified markers on the omnigrid mat. That includes two poses on the z plane of the mat (z=0cm) and three poses with thumb, index or both fingertips on cubes of known dimensions(z=2.54cm). At each pose we recorded a rosbag file with the raw RGB image, depth point cloud, camera info and TF tree [9] for a duration of 9.7 seconds, i.e. 30 frames per second (fps) that corresponds to approximately 290 messages for each topic.

Poses A-B, C-D, E-F, F-G and F-H were used to evaluate the *accuracy* of our method with respect to ground truth positions. In these poses fingertip positions were fairly distinct from each other, e.g. aperture of 14cm. To assess the discriminability of the proposed system, i.e the minimum difference that can be detected between apertures, poses I1, I2, I3, I4 and I5 were used.

### 4.3 3D Hand Keypoints detection

Our system utilizes 2D hand poses estimated from OpenPose [3, 4, 20, 22] and depth information in the form of point cloud to extract 3D hand keypoints[3]. We used the Robotic Operating System

[1] http://wiki.ros.org/ar_track_alvar
[2] https://github.com/Roboskel-Manipulation/manos_vision
[3] https://github.com/Roboskel-Manipulation/openpose_utils

(ROS) [19] to build the whole pipeline as depicted in Figure 3. The OpenPose's real-time processing performance, that is the number of processed frames per second, is constrained by the performance of the GPU available. In this work, we aimed at evaluating our method's performance at all collected frames. For this reason we built a ROS service that reads camera frames on demand from a rosbag file. In our case, this occurs when 3D keypoint extraction from a given frame is completed (BagServer and Bag Client in Figure3). The OpenPose ROS wrapper subscribes on the raw RGB

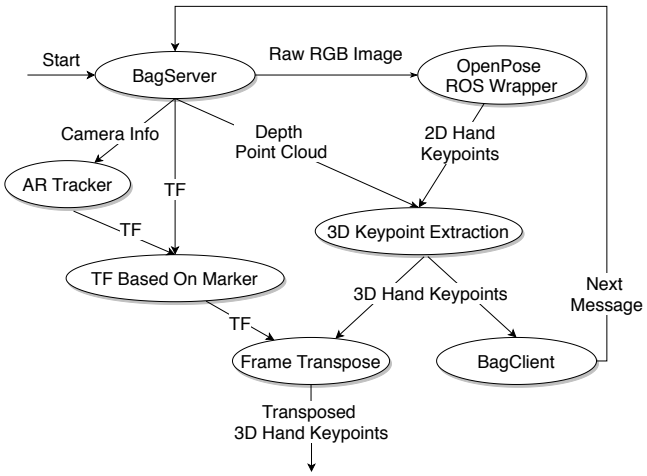



**Figure 3: ROS Node Graph**

image topic to estimate the 2D hand keypoints from a human pose (Figure 4). Even though the OpenPose system provides information for keypoints throughout the whole body, we only focus on the hand keypoints and especially the right-hand index and thumb tips. The 3D Pose Extraction node uses the information from Hand 2D Poses topic and combines it with the corresponded messages on the Depth Point Cloud topic published on demand too. This node publishes messages on the 3D Hand Poses topic and the Frame Transpose node transforms them onto the frame of the AR marker placed on the workspace. To achieve that, it uses the TF tree provided by the TF Based On AR Marker node. Each time that a message is being published on the 3D Hand Pose topic node, the BagClient node asks from the BagServer to read and publish the next message for every topic in the rosbag file examined.

## 5 EVALUATION

### 5.1 Accuracy Assessment

Overall, the mean Euclidean distances are below 2 cm for most finger positions, except F(G) and F(H) (Table 1). Moreover, the standard deviation of the mean Euclidean distance is at most 0.29 cm. This shows that the measurements of the fingertip are compact and consistent. In Table 1 we present the mean value (and the standard deviation) of the absolute difference between the ground truth and the measured coordinates for every fingertip position. In Figures 5 and 6 we show an example of the ground truth points and the detected fingertip positions at pose A-B (for index and thumb respectively for x-y, x-z and y-z coordinate frames).

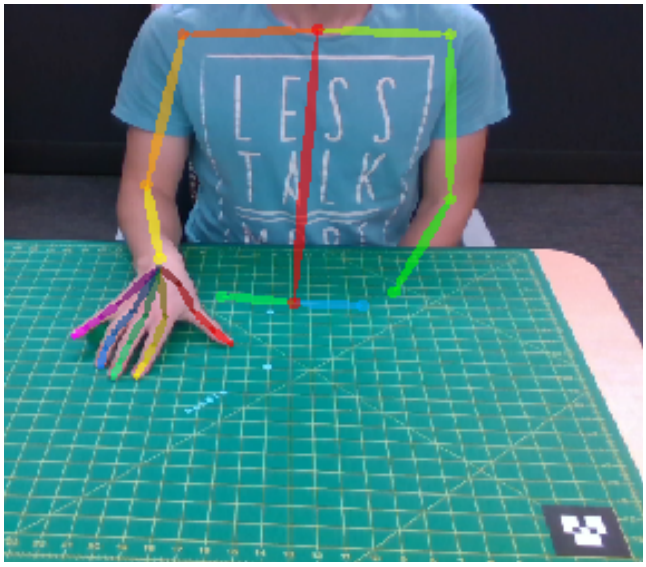


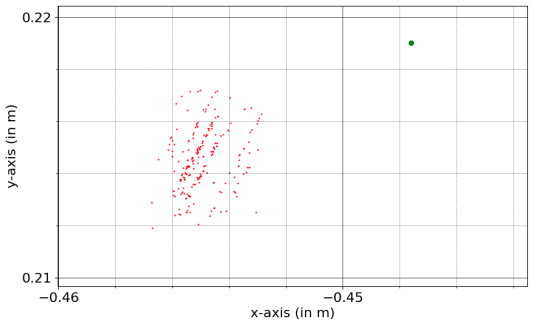
**Figure 4: Example of 2D human poses acquired from OpenPose ROS wrapper.**

| Ground Truth (GT) Positions (x, y, z) | Mean Measured Difference from GT | | | Mean Euclidean Distance (cm) |
|---|---|---|---|---|
| | X(cm) | Y(cm) | Z(cm) | |
| **A** (34.60, 29.52,0.0) | 0.15(0.10) | 1.01(0.17) | 00.89(0.18) | **1.24**(0.18) |
| **B** (44.76, 21.90, 0.0) | 0.73(0.07) | 0.45(0.11) | 0.87(0.18) | **1.36**(0.18) |
| **C** (29.52, 37.14, 0.0) | 0.22(0.12) | 1.71(0.18) | 1.06(0.16) | **1.1**6(0.17) |
| **D** (29.52, 24.40, 0.0) | 0.14(0.10) | 0.94(0.16) | 0.65(0.15) | **2.03**(0.17) |
| **E** (40.03, 34.00, 0.0) | 0.61(0.17) | 1.65(0.20) | 01.57(0.22) | **1.46**(0.27) |
| **F**(E) (43.74, 30.50, 2.54) | 0.02(0.16) | 0.54(0.27) | 1.34(0.25) | **1.92**(0.27) |
| **F**(G) | 0.41(0.09) | 0.02(0.19) | 1.53(0.25) | **2.91**(0.26) |
| **F**(H) | 0.93(0.15) | 0.18(0.23) | 1.62(0.27) | **2.31**(0.26) |
| **G** (47.55, 26.23, 0.0) | 0.01(0.07) | 1.78(0.12) | 02.71(0.25) | **1.82**(0.29) |
| **H** (51.36, 29.77, 2.54) | 0.40(0.17) | 0.24(0.27) | 2.23(0.30) | **1.95**(0.29) |

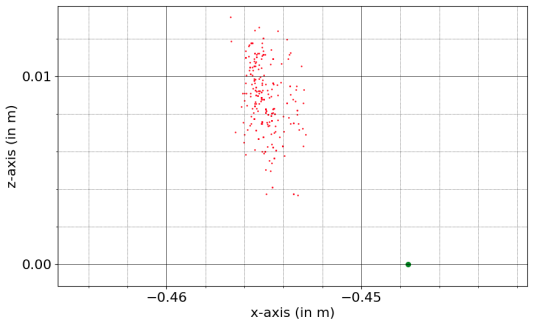
**Table 1: Absolute difference of mean values of detected fingertip coordinates and their mean Euclidean distance from ground truth in cm of measured fingertips. The ground truth stickers position are under the point name. The Standard deviation is in parenthesis next to each mean.**

By examining each axis separately, we see that the mean values of x-axis present an absolute difference of less than 0.93 cm (point F(H)). The highest errors are observed on the y-axis (1.78 cm for point G) and the z-axis (2.71 cm for point G). These higher mean differences are related to errors introduced by the depth information especially on the z-axis (depth) . Such systematic errors could

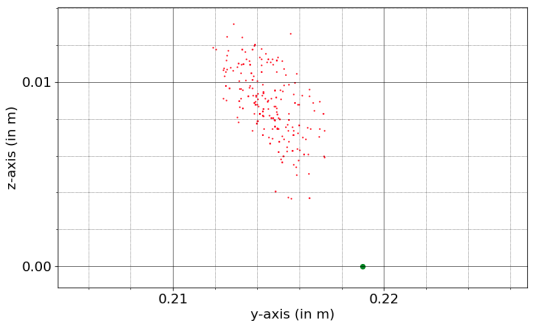
be corrected by calibrating the system. Finally, all measurements present a low standard deviation that does not exceed 0.3 cm on any axis, indicating a high repeatability of OpenPose results.




**(a) X-Y**




**(b) X-Z**

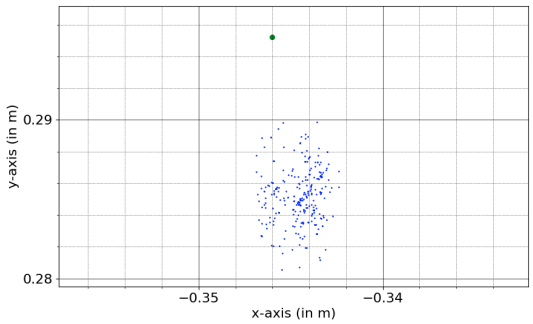



**(c) Y-Z**

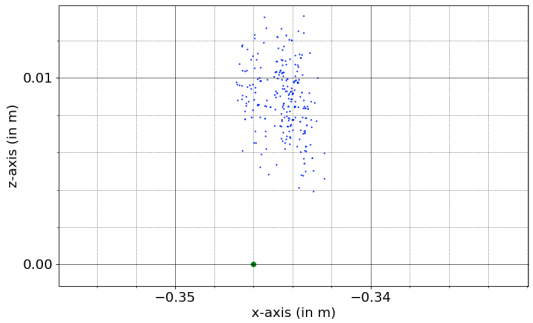
**Figure 5: Index at AB Pose. Green Point is the Ground truth sticker.**

| Ground Truth | Measured Means |
|---|---|
| 1cm | 0.6cm(0.1) |
| 2cm | 1.4cm(0.1) |
| 3cm | 2.5cm(0.1) |
| 4cm | 3.2cm(0.1) |
| 5cm | 4.3cm(0.1) |

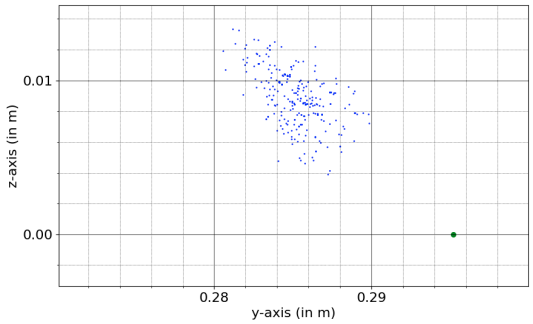
**Table 2: Mean values of Measured Apertures. The standard deviation is in parenthesis.**




**(a) X-Y**




**(b) X-Z**




**(c) Y-Z**

**Figure 6: Thumb at AB Pose. Green Point is the Ground truth sticker.**

## 5.2 Discriminability

The proposed system exhibits excellent discrimination capabilities even in the extreme case of 1cm aperture, where the two fingers are in touch with each other. Due to OpenPose's excellent performance and camera's accurate depth information, there is no overlapping between the detected fingertip positions on any plane. This is shown in the box plot of Figure 7, where we do not observe extreme outliers in any aperture size. Overall, there is a systematic underestimation of the apertures ranging on average from 0.4 to 0.8 cm (Table 2). However, there is no overlap between the box-plots indicating a robust discrimininability of the aperture between the two fingers. The average difference between the measured apertures is relatively constant and very close to the ground truth one (1cm), e.g. the difference of the measured apertures between 1cm and 2cm is 0.86

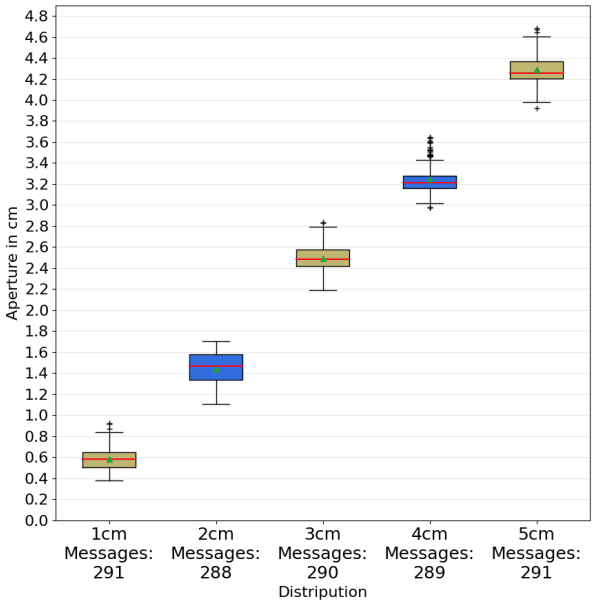



**Figure 7: Boxplot for 1cm-5cm fingertips aperture. The median value is denoted as a red line inside the box and the mean value with a green triangle. Outliers are denoted with crosses.**

cm, between 2 cm and 3 cm is 1.05 cm, between 3 cm and 4 cm is 0.75 cm and between 4 cm and 5 cm is 1.04 cm.

## 6 DISCUSSION

In this paper, we evaluated a flexible and accessible method to detect 3D markerless hand poses for a wide spectrum of applications using a commercial low-cost RGB-D camera. The 3D hand poses were estimated with the integration of OpenPose, a state-of-the-art 2D hand pose estimator , and depth information acquired from the camera. We evaluated the system's accuracy and discriminability examining 10 different static poses. The estimation of the 3D hand poses did not exhibit errors greater than 2.71 cm in any axis., while the standard deviation of the measurements was below 0.3cm, indicating high repeatability of the methods. Finally, we show the very good discriminability of the proposed system, even for the extreme case of the two fingertips touching each other.

The deviations from ground truth position can be attributed to errors introduced in all stages of the pipeline. First of all, there is a small bias in the measurements from the placement of the user's fingers in the markers; naturally, the fingers occupy a certain volume and are related to a neighbourhood of points around ground truth positions. In addition, Intel RealSense D435i camera[4] introduces an error less than 2% for up to 2 meters and 80 percent field of view. Thus, we should expect a small error on x and y axes from finger misplacement and a distributed depth error from the camera on y and z axes. On the top of that, we should take into consideration the small error that the OpenPose[20] introduces on x and y axes, as well as the transformation error from the camera's frame to the frame of the AR marker, due to the marker's shape distortion.

The systematic errors resulting from camera characteristics and coordinate transformation can be alleviated by taking into account the evaluation presented in this work. In the future we intend to collect data from more participants and extend this work to hand movements estimation.

To the best of our knowledge, there are no other tools integrated into ROS, except OpenNi[5] that can be used for tracking human hands using RGB-D sensors. In the future it would be interesting to compare the performance of OpenNi and to integrate more state-of-the-art methods in ROS and compare their performance.

## ACKNOWLEDGMENTS

This work was supported in part by National Science Foundation under award number 1719031. This work was also supported in part by the "Stavros Niarchos Foundation" Industrial Post-doc Fellowship of NCSR "Demokritos" on Human-Robot Collaboration: human collaborator representation for robot autonomous decisions, Roboskel lab, the robotics activity of Software and Knowledge Engineering Lab, Institute of Informatics and Telecommunications, NCSR "Demokritos".

## REFERENCES

[1] Caterina Ansuini, Andrea Cavallo, Atesh Koul, Marco Jacono, Yuan Yang, and Cristina Becchio. 2015. Predicting Object Size from Hand Kinematics: A Temporal Perspective. *PLOS ONE* 10 (03 2015), 1–13. https://doi.org/10.1371/journal.pone.0120432

[2] Miguel Arduengo, S Jorgensen, K Hambuchen, L Sentis, F Moreno-Noguer, and G Alenya. 2017. Ros wrapper for real-time multi-person pose estimation with a single camera. *Institut de Robotica i Informatica Industrial, CSIC-UPC, Tech. Rep* (2017).

[3] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008.*

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR.*

[5] Andrea Cavallo, Atesh Koul, Caterina Ansuini, Francesca Capozzi, and Cristina Becchio. 2016. Decoding intentions from movement kinematics. *Scientific Reports* 6 (11 2016), 37036. https://doi.org/10.1038/srep37036

[6] B. Chandrasekaran and J. M. Conrad. 2015. Human-robot collaboration: A survey. In *SoutheastCon 2015.* 1–8. https://doi.org/10.1109/SECON.2015.7132964

[7] Ali Erol, George Bebis, Mircea Nicolescu, Richard D. Boyle, and Xander Twombly. 2007. Vision-based hand pose estimation: A review. *Computer Vision and Image Understanding* 108, 1 (2007), 52 – 73. https://doi.org/10.1016/j.cviu.2006.10.012 Special Issue on Vision for Human-Computer Interaction.

[8] Terrence Fong, Charles Thorpe, and Charles Baur. 2003. Collaboration, Dialogue, Human-Robot Interaction. In *Robotics Research*, Raymond Austin Jarvis and Alexander Zelinsky (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 255–266.

[9] Tully Foote. 2013. tf: The transform library. In *Technologies for Practical Robot Applications (TePRA), 2013 IEEE International Conference on (Open-Source Software workshop).* 1–6. https://doi.org/10.1109/TePRA.2013.6556373

[10] Elena Grigore, Kerstin Eder, Alexander Lenz, Sergey Skachek, Anthony Pipe, and Chris Melhuish. 2011. Towards Safe Human-Robot Interaction, Vol. 6856. 323–335. https://doi.org/10.1007/978-3-642-23232-9_29

[11] Shangchen Han, Beibei Liu, Robert Wang, Yuting Ye, Christopher Twigg, and Kenrick Kin. 2018. Online optical marker-based hand tracking with deep labels. *ACM Transactions on Graphics* 37 (07 2018), 1–10. https://doi.org/10.1145/3197517.3201399

[12] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. 2018. Hand Pose Estimation via Latent 2.5D Heatmap Regression. In *The European Conference on Computer Vision (ECCV).*

[13] Przemyslaw Lasota, Terrence Fong, and Julie Shah. 2017. A Survey of Methods for Safe Human-Robot Interaction. *Foundations and Trends in Robotics* 5 (01 2017), 261–349. https://doi.org/10.1561/2300000052

[4] https://www.intelrealsense.com/depth-camera-d435i/

[5] http://wiki.ros.org/openni

[14] Xiu Li, Zhen Fan, Yebin Liu, Yipeng Li, and Qionghai Dai. 2019. 3D Pose Detection of Closely Interactive Humans Using Multi-View Cameras. In *Sensors*.

[15] Charles Malleson, John P. Collomosse, and Adrian Hilton. 2019. Real-Time Multi-person Motion Capture from Multi-view Video and IMUs. *International Journal of Computer Vision* (2019), 1 – 18.

[16] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. 2018. GANerated Hands for Real-Time 3D Hand Tracking from Monocular RGB. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*. 11. https://handtracker.mpi-inf.mpg.de/projects/GANeratedHands/

[17] Nobuyasu Nakano, Tetsuro Sakura, Kazuhiro Ueda, Leon Omura, Arata Kimura, Yoichi Iino, Senshi Fukashiro, and Shinsuke Yoshioka. 2019. Evaluation of 3D markerless motion capture accuracy using OpenPose with multiple video cameras. *bioRxiv* (2019). https://doi.org/10.1101/842492 arXiv:https://www.biorxiv.org/content/early/2019/11/15/842492.full.pdf

[18] P. Panteleris, I. Oikonomidis, and A. Argyros. 2018. Using a Single RGB Frame for Real Time 3D Hand Pose Estimation in the Wild. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 436–445. https://doi.org/10.1109/WACV.2018.00054

[19] Morgan Quigley, Ken Conley, Brian P. Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob C. Wheeler, and Andrew Y. Ng. 2009. ROS: an open-source Robot Operating System. In *ICRA 2009*.

[20] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. 2017. Hand Keypoint Detection in Single Images using Multiview Bootstrapping. In *CVPR*.

[21] J. T. C. Tan and T. Arai. 2011. Triple stereo vision system for safety monitoring of human-robot collaboration in cellular manufacturing. In *2011 IEEE International Symposium on Assembly and Manufacturing (ISAM)*. 1–6. https://doi.org/10.1109/ISAM.2011.5942335

[22] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional pose machines. In *CVPR*.

[23] H. A. Yanco and J. Drury. 2004. Classifying human-robot interaction: an updated taxonomy. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, Vol. 3. 2841–2846 vol.3. https://doi.org/10.1109/ICSMC.2004.1400763

[24] Christian Zimmermann and Thomas Brox. 2017. Learning to Estimate 3D Hand Pose From Single RGB Images. In *The IEEE International Conference on Computer Vision (ICCV)*.