




Combining Static and Dynamic Predictions of Transfer Points for Human Initiated Handovers

Janneke Simmering¹, Sebastian Meyer zu Borgsen^{1,2}(✉) , Sven Wachsmuth^{1,2}, and Ayoub Al-Hamadi³

¹ Bielefeld University, 33594 Bielefeld, Germany

{jsimmering, semeyerz, swachsmu}@techfak.uni-bielefeld.de

² CITEC, Bielefeld University, 33594 Bielefeld, Germany

³ IIKT, University of Magdeburg, Magdeburg, Germany
ayoub.al-hamadi@ovgu.de

<https://www.cit-ec.de>, <http://www.iikt.ovgu.de/nit.html>

Abstract. In many scenarios where robots could assist humans, handover situations are essential. But they are still challenging for robots, especially if these are initiated by the human interaction partner. Human-human handover studies report average reaction times of 0.4 s, which is only achievable for robots, if they are able to predict the object transfer point (OTP) sufficiently early and then adapt to the human movement. In this paper, we propose a hand tracking system that can be used in the context of human initiated handover as a basis for human reaching motion prediction. The OTP prediction implemented is based on the minimum jerk model and combines a static estimation utilizing the human's initial pose and a dynamic estimation from the current hand trajectory. Results are generated and analyzed for a broad spectrum of human initiated scenarios. For these cases we examine the dynamics of different variants of the proposed prediction algorithm, i.e., how early is a robot's prediction of the OTP within a certain error range? The tracking delivers results with an average delay, after the initialization, of 0.07 s. We show that the OTP prediction delivers results after 75 % of the movement within a 10 cm precision box.

1 Introduction

Object handovers take place everywhere in our daily lives. With robots becoming more and more present in our common working or living environment, this is an essential part of the socially accepted interaction with them. Such close interactions demand a collaborative and precise synchronization of both partners in

S. Meyer zu Borgsen and S. Wachsmuth—This research/work was supported by the Cluster of Excellence Cognitive Interaction Technology ‘CITEC’ (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG).

© Springer Nature Switzerland AG 2019

M. A. Salichs et al. (Eds.): ICSR 2019, LNAI 11876, pp. 676–686, 2019.

https://doi.org/10.1007/978-3-030-35888-4_63

space and time. Previous handover studies provided evidence that an improved reactivity of the robot increases the acceptance of such systems [11]. This requires an early prediction of the object transfer point (OTP), which defines the object's position in 3d space where the actual object transfer event takes place (see marked coordinate system in Fig. 1(b)). This point can be inferred from the tracking of the hand position. Existing approaches can be distinguished by several factors. These include the number of cameras used, whether color, depth images or a combination of both is used and if the output is only one point describing the hand center position or if the hand's articulation is tracked. The proposed system should enable a robot to react to a handover considering a human-like timing. From human-human handover studies an average handover duration of about 1.24 s was reported [1]. The average reaction time was found to be about 0.4 s [12]. Therefore, the processing of the hand tracking and the planning of the robot's motion should only induce a minimal delay below 0.4 s. Approaches with a high processing time (e.g., Bray et al. [2]) do not fulfill these requirements. Hackenberg et al. [6] report a rate of more than 50 fps and rely only on depth images. But they assume strong visibility constraints, e.g., the palm always facing the camera with visible fingers, which are not always met in handover scenarios with occluding objects. Predicting human reaching motions in a handover scenario is a very recent research topic. OTP predictions have been tackled by [7, 8, 14]. Most approaches learn human motions and need many human demonstrations or use imitation learning [10, 13]. In order to generalize different handover scenarios and individuals large data sets are needed. Li and Hauser present a number of different models to predict human reaching motions, but only present preliminary results for two of them [9]. Nemlekar et al. combine a modified version of the Probabilistic Movement Primitives approach from Maeda et al. with an offline component in order to predict an estimated OTP as soon as a handover intent is detected [10, 12]. Our proposed system is based on Nemlekar's approach but substitutes the Probabilistic Movement Primitives component with the minimum jerk model, which is the first model presented by Li and Hauser [9]. The minimum jerk model was originally proposed by Flash and Hogan as a mathematical model for describing human arm movements [5]. Our approach is inspired by Chen et al. [4] who report a rate of more than 300 fps. As the robot should interact autonomously, our method relies only on one depth camera which is mounted in the robot's chest. Similar to Chen et al., we segment the hand and estimate the hand center based on the boundary points of the segmented region providing an input for a mean-shift based tracking. For the OTP prediction, we apply a model-based approach utilizing the minimum jerk model combined with the basic idea of Nemlekar et al. [5, 9, 12]. Instead of optimizing the approach for a minimal displacement at the end of the tracking, we examine the effects of different combination schemes of static and dynamic OTP prediction in the early reaching phase.

2 Approach

Similar to Nemlekar et al. [12], our system combines an offline and online component for the OTP prediction, but we examine the combination schemes in a more sophisticated manner. The minimum jerk model is a simple method for trajectory generation minimizing a cost criterion based on the trajectory’s jerk [5]. It is used in the online component and predicts the human hand motions based on the current hand position, the handover start time, and the estimated end time [9]. The current hand position is provided by a hand tracking system (Sect. 2.1) that is used to continuously update the OTP prediction. The initialization of the tracking system is given by a hand detection on a color image based on OpenPose [3]. This provides accurate results but takes up some processing time which leads to an initial delay of the tracking results of between 0.3 s and 0.4 s. This 2D position is then converted, based on the corresponding depth image, into a 3D position which is further used for low-latency tracking. Point clouds are buffered in order to deal with the initial delay and to provide a continuous tracking input. The hand tracking relies only on point clouds. The camera used is an Intel RealSense D435, which has a wider opening angle for the depth sensor than for the color image and, therefore supports a larger interaction space during the tracking after the initialization. The OTP is defined as the 3D position at which the object will be handed over. With the approach from Nemlekar et al. [12] a prediction for this point is available as soon as the intent for a handover is detected.

2.1 Hand Tracking with Point Clouds

The hand tracking system, adapted from Chen et al. [4], purely relies on point clouds. It determines the hand position (x_{hand}) in the current frame, as well as the current velocity of the hand, given the hand position from the previous frame. In each frame, the point x closest to the previous hand position, in the point cloud, is found as a seed for the current hand hypothesis. Chen et al. use the geodesic distance of points to avoid including other body parts in the segmented hand region [4]. A similar effect can be achieved by removing points from the point cloud whose distance to x is greater than a certain distance. This region is used as hand region (r_{hand}). The points at the edge of the region r_{hand} are found as boundary points. The hand position is then approximated as the point (x_{approx}) with the most boundary points within a radius d_{ms} , i.e., assuming it at the end of the arm [4]. A mean-shift step then refines the hand position [4]:

$$x_{hand} = \frac{\sum_{p \in r_{hand}} p * 1_{d(p, x_{approx}) < d_{ms}}}{\sum_{p \in r_{hand}} 1_{d(p, x_{approx}) < d_{ms}}} \quad (1)$$

The calculated position x_{hand} is the result of the current hand tracking and serves as the input for the next tracking step.

2.2 Object Transfer Point Prediction

For the OTP prediction a static and a dynamic OTP are used (see [12]). The weighted sum of these two points, the integrated OTP, is the actual output that is sent to the robot as a grasp position goal. During the handover the static OTP is calculated once based on the initial position of the human partner as half-way between robot and human. The dynamic OTP is calculated based on the model defined below utilizing the reaching motion of the human partner, so far observed. The static OTP is instantly available at the start of a handover, so that the robot can start to reach out as soon as the handover intent is detected. On the one hand, it is also a relatively safe prediction, that typically is not too far away from the actual OTP. Therefore, it can reduce the prediction error that the dynamic OTP initially might have because it is missing a sufficient motion history. On the other hand, the dynamic OTP can adapt the prediction to the current situation and is supposed to provide a more accurate position than the static OTP, once a certain amount of the movement was observed. The difference in the movement speed of humans and robots allows the dynamic OTP to become more accurate early during the robot's movement, because the human's motion will proceed faster than the robot's motion. Thus, the key objective is to weight these OTP predictions accordingly, so that a robot is able to react fast and accurately to human initiated handovers.

The utilization of the minimum jerk model for the dynamic OTP prediction is based on the assumption that human motions are smooth, which means, that the jerk of the motion is as small as possible. To predict the OTP, the end time of the movement is needed in order to resolve the minimum-jerk equation for the end position x_f [5]. Here an average handover time can be used. Basili et al. found that the time between the initiation of the handover and the actual handover is $1.24\text{ s} \pm 0.28\text{ s}$ [1]. Due to the bell shaped velocity profile the end time can be updated during the hand over at the velocity extremum, which can be detected. It is reached when 50 % of the complete movement is executed. Consequently, the end time can be predicted at this point by adding the time that already passed since the handover started ($t_c - t_0$) to the current time. Thus, the minimum jerk model for x_f , which is the predicted end position of the movement and therefore the dynamic OTP, is:

$$x_f = \frac{x_c - x_0}{10 * \left(\frac{t_c - t_0}{t_f - t_0}\right)^3 - 15 * \left(\frac{t_c - t_0}{t_f - t_0}\right)^4 + 6 * \left(\frac{t_c - t_0}{t_f - t_0}\right)^5} + x_0 \quad (2)$$

The error of the model should decrease over the observation time.

The static and dynamic OTP are interpolated to calculate the integrated OTP as the result of the prediction. This interpolation can either be applied to the whole time period, or only when the dynamic OTP is reasonably stable. Otherwise, the static OTP can be used solely, until the dynamic OTP is continuously in a reasonable range, i.e. within the space between human and robot. For the weight calculation there are two possible functions. It can either be calculated as ratio of the movement executed – where t_s is the time at which

the dynamic OTP delivers usable results – or with a parable function that has its maximum in the origin and goes through the point (1,1) [12]:

$$W_{lin} = \frac{t_c - t_s}{t_f - t_s}, \quad W_{quad} = 0.2 * \left(\frac{t_c - t_s}{t_f - t_s}\right) + 0.8 * \left(\frac{t_c - t_s}{t_f - t_s}\right)^2 \quad (3)$$

The integrated OTP is calculated as formulated in Eq. 4 [12].

$$OTP_{integrated} = W * OTP_{dynamic} + (1 - W) * OTP_{static} \quad (4)$$

In the context of mobile robots that autonomously perform safe movements, there are some traits the prediction should have. First the overall error of the prediction should be as small as possible to avoid that the robot moves in unpredictable directions which would confuse the human partner. The prediction should also be as good as possible early during the handover and be able to deal with different users. We implemented three different features that are supposed to provide these traits and can be combined for an improved prediction:

Feature 1 (Initial fixation). *Fixate static prediction until the dynamic prediction reaches the interaction space estimated.*

Feature 2 (Linear interpolation). *Use linear interpolation instead of quadratic.*

Feature 3 (Update end time). *Update end time after peak velocity detection (extremum of bell shaped profile).*

3 Evaluation

In the evaluation we assess the precision of the estimated OTP over time and analyze the influence of the integration scheme (Eq. 4) including Features 1–3. Therefore, we recorded several handover scenarios from the robot’s perspective (see Fig. 1(c)–(e)). This data was used to evaluate the success rate of the hand tracking, the time delay of the tracking data, as well as the quality of the OTP prediction.

Test Cases and Evaluation Procedure: In the test cases considered, we systematically vary several aspects of handover configurations in order to test the generality of the model: (i) initial position relative to the robot, (ii) giver vs. receiver, (iii) small vs. large objects, (iv) direct handover vs. pre-grasp/hand change, (v) normal vs. tall person, (vi) posture with hands closer vs. farther away from the body. The general process of the test cases is given in Fig. 1. The human interaction partner approaches the robot from a distance of about 2 m and moves his/her arm to the handover position in front of the robot (see coordinate system in Fig. 1(b)). The case where the human stands straight in front of the robot and acts as a receiver was recorded seven times in total. Five of these recordings are from a person that is 1.7 m tall and, additionally, vary the postures by keeping the hands closer or farther away from the body. The other two are conducted

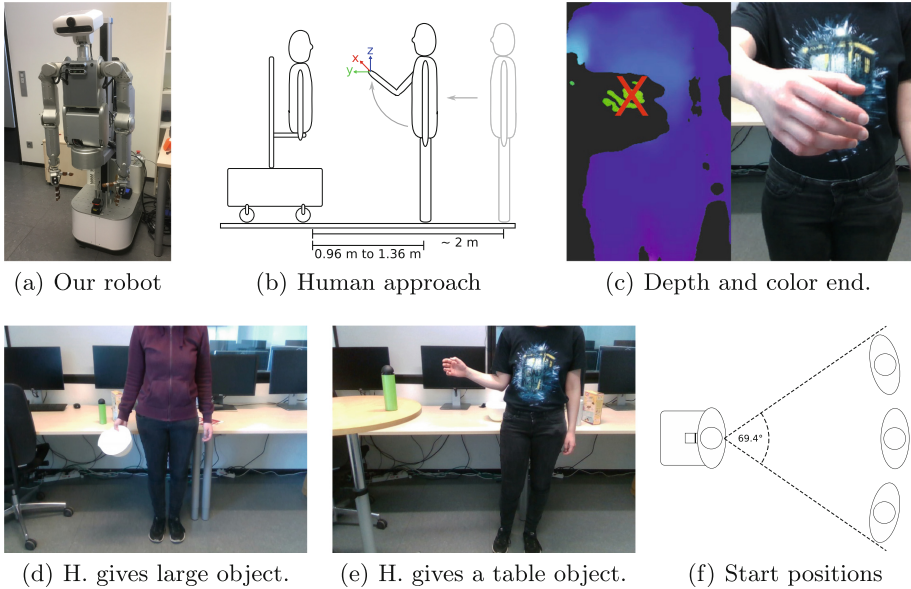


Fig. 1. Experiment Setup: A participant receives or gives an object from/to the Meka M1 Mobile Manipulator robot (a). Starting 2 m away the human approaches the robot and reaches the object transfer point (coordinate system in (b), red X in (f)). The final position in the interaction with the annotation of the reference position is shown as depth and color image (f). Start position (c) and start configurations (d, e) are varied. (Color figure online)

with a person who is 1.9 m tall. A similar situation with the human acting as giver was recorded for four different objects (two smaller bottles and two bulky objects, a bowl and a box of cereals). In another case one of the bottles was held with both hands in the beginning; in three cases the object was moved into the hand before the handover either by taking the object from a table next to the human or by passing it from the left to the right hand. The start position was varied in three cases where the human stands at the edge of the robot's field of view; in one of them the human acted as giver in the other two as receiver. The last scenario recorded a longer interaction in which the human approached the robot and waved his/her right hand around before performing a reaching motion. Altogether, 19 recordings of handover scenarios were used for the evaluation.

For each scenario, the camera stream of the robot and the tracking result was recorded; the reference handover position was manually annotated (see Fig. 1(c)). Then, the OTP prediction was performed multiple times on the same hand-tracking data using different configurations of activated algorithmic features: *initial fixation* (1), *linear vs. quadratic interpolation* (2), and *update end-time* (3). The prediction starts once a handover intent is detected which is defined as significant increase of the velocity towards the robot while the velocity to the left and right and height are relatively small. The OTP is continuously predicted

until the estimated end time of the expected OTP is reached or the velocity data shows a movement of the hand away from the robot. For each OTP prediction, we measure the difference to the annotated OTP in x-/y-/z-direction. The reference frame for the evaluation is defined as the annotated handover position with the z-axis pointing upwards, the y-axis describing the distance between the human and the robot, and the x-axis along left and right (cf. Fig. 1(b)). Thus, even for the end point of the tracked hand movement, there can be a deviation from the reference position because the tracking result may not be precise (e.g. occlusion by bulky transfer-objects) or the end time may have been incorrectly estimated.

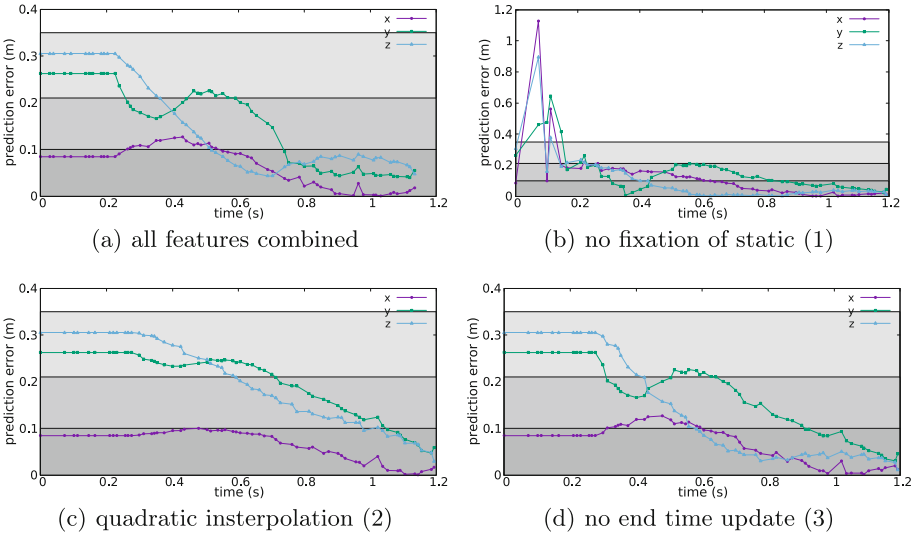


Fig. 2. Absolute error of the OTP prediction over time for four different feature combinations on the example of a vis-à-vis handover as the most common example. Error ranges that contain most test cases when all features are activated: light gray 0.35 m for overall error y and z direction, gray 0.2 m for overall error x direction and dark gray 0.1 m end time error range for all directions

Overall Results: In Fig. 2 results of different system configurations are shown. It can be seen that (i) without fixation (Fig. 2(b)) the OTP prediction in the first 0.2s is unstable; (ii) with quadratic interpolation (Fig. 2(c)) the prediction error shrinks slower; a similar effect shows up (iii) without end time update (Fig. 2(d)). Thus, each algorithmic feature enhances the prediction either by reducing the overall error or by reaching a smaller deviation earlier in time. The initialization of the tracking, i.e. the initial hand pose estimation on the rgb frame, takes between 0.3s and 0.4s. During this phase, the system buffers about 10 depth frames, which are processed afterwards by the tracking. In the following, this delay is decreased to 0.07s on average leading to very short reaction times of the robot (see Fig. 3).

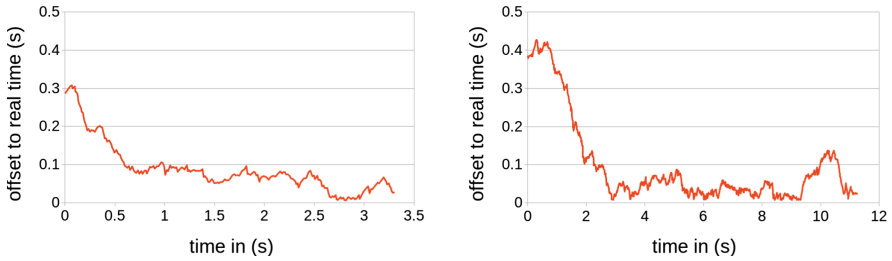


Fig. 3. Delay of the hand tracking for a normal and the long test case.

In order to provide an evaluation over all test cases, we define different error ranges and report how many test cases are within these error ranges for different system configurations. First we measure the error ranges of the configuration with all features activated. The maximal error in 17 of 19 test cases (over the complete time interval) is below 35 cm for the y- and z-axes and in 18 of 19 test cases it is below 20 cm for the x-axis. Thus, we define $[20\text{ cm} \times 35\text{ cm} \times 35\text{ cm}]$ as a first coarse precision box. Considering the *end-time error* at the detected OTP, 11 of the 19 test cases are below 10 cm in $[x/y/z]$. This defines a second more restrictive precision box (the error ranges of both precision boxes – the coarse: $[20 \times 35 \times 35]$ and fine: $[10 \times 10 \times 10]$ – are visualized in Fig. 2). If the fine precision box is relaxed from a 10 cm to a 15 cm range, 15 of 19 cases are finally within it. Ideally, the OTP prediction should stay in the coarse precision box during the complete movement. Here a low deviation along the x-axis ($<20\text{ cm}$) is important because a larger deviation would trigger the robot to turn away. In the different test cases, we noticed a single case where the x-corridor was slightly missed (max. x-error 20.4 cm). This has been one of cases where the human initially stands at the edge of the robot’s field of view. A larger deviation along the y-axis would trigger the robot to adjust to a wrong distance to the human. This has been missed only two times, where the end-point has been wrongly estimated.

Considering the influence of the algorithmic features, we measured that only 13 instead of 17 cases stay inside the coarse precision box $[20 \times 35 \times 35]$ along the y- and z-axes, if no end time update is performed. On the downside, in two cases an update of the end-time lead to a slight miss of the $10 \times 10 \times 10$ precision box. In the case where all features are activated, the OTP predictions converge to the fine precision box ($<10\text{ cm}$) after 75% of the movement. This increases to 82%, if there is no end-time update. The 15 cm-precision box converges after 64% of the movement, if all algorithmic features are activated. For the quadratic interpolation, the OTP prediction reaches the fine 10 cm-precision box after 90% of the movement in only 9 of 19 test cases.

4 Discussion

Human-initiated handovers are a frequent case in human-robot interaction. It should work for different people, with different objects, from different directions, and even if the human does a preparation movement beforehand. Therefore, we selected a combination of simplified models considering the body pose and orientation as well as the dynamic movement of the hand, that both work without any pre-learning of motion models. In the evaluation, we show that even if the final OTP predictions (end-time error after hand tracking) may be similar, there are large differences between OTP predictions during the movement of the human's hand. These early predictions are most important, when the robot should smoothly react on the human-initiated handover. However, these are rarely analyzed systematically. Thus, we take a deeper look at the impact of different algorithmic combination schemes on these prediction dynamics. The results show that the motion-based OTP prediction is quite poor in the beginning. Therefore, Feature 1 (Initial fixation) improves the prediction by limiting the OTP prediction error. This leads to a smoother interaction with the robot, because its movements are more predictable and human-like. The quadratic interpolation does not exploit the fast improvement of the dynamic prediction, thus the end error range is reached for fewer cases and later during the handover. The end-time update (Feature 3) together with the dynamic prediction is responsible for adjusting the prediction to the partially observed handover. This manifests itself in an earlier convergence to the fine precision box defined by the error range for the end-time. Overall, the features described in Sect. 2.2 enhance the prediction results and reach a smaller error range early during the handover while maintaining a limited maximal error over the complete movement. This fulfills the requirement, that the robot is able to move in the right direction early during a handover.

Limitations of the system were found for large objects, like bowls. These were already hard to see for the depth sensor due to shadows which resulted in a poor hand tracking and prediction since in this case the end error range was not reached in any of the directions. Furthermore, seven other cases did not reach the end-time error range in at least one direction, which is mostly due to the noisy velocity data which can lead to an early termination of the prediction process, because the end of the handover movement is wrongly detected. Therefore, the method still can be improved by less noisy velocity data and more advanced detection methods for the start and end-time of the handover.

5 Conclusion

We presented an object transfer point (OTP) prediction which combines a static and a dynamic prediction of an OTP for human initiated object handover from and to a robot. This approach is integrated with our low-latency handtracking algorithm, for which we combined state-of-the-art techniques to estimate the position of a human hand in a color image and track it on depth clouds. This

way we achieved a tracking performance of on average 0.07s delay after the initialization phase. The whole system was able to run on off-the-shelf hardware on a mobile humanoid service robot. We could show that the combination of a static and dynamic OTP for predicting the handover position is useful for an early estimation that becomes more accurate over time, allowing the robot to start its movement early and adapt over time, decreasing wait times and allowing precise adaption to the human interaction partner. As we build upon the minimum-jerk-model, we do not need to train the system with lots of data and are independent of retraining for a different kind of robot. Evaluation on a wide range of situations showed validity of this approach. This leads to robots that adapt to the human and, thus facilitate socially accepted human-robot interactions.

References

1. Basili, P., Huber, M., Brandt, T., Hirche, S., Glasauer, S.: Investigating human-human approach and hand-over. In: Vernon, D., et al. (eds.) *Human Centered Robot Systems*, vol. 6. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-10403-9_16
2. Bray, M., Koller-Meier, E., Van Gool, L.: Smart particle filtering for 3D hand tracking. In: *Proceedings of Sixth IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, Seoul, Korea (2004)
3. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. [arXiv:1812.08008](https://arxiv.org/abs/1812.08008) (2018)
4. Chen, C., Chen, Y., Lee, P., Tsai, Y., Lei, S.: Real-time hand tracking on depth images. In: *2011 Visual Communications and Image Processing (VCIP)*, November 2011
5. Flash, T., Hogan, N.: The coordination of arm movements: an experimentally confirmed mathematical model. *J. Neurosci.* **5**(7), 1688–1703 (1985)
6. Hackenberg, G., McCall, R., Broll, W.: Lightweight palm and finger tracking for real-time 3D gesture control. In: *2011 IEEE Virtual Reality Conference*, March 2011
7. Huber, M., et al.: Evaluation of a novel biologically inspired trajectory generator in human-robot interaction. In: *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*, September 2009
8. Kajikawa, S., Ishikawa, E.: Trajectory planning for hand-over between human and robot. In: *Proceedings 9th IEEE International Workshop on Robot and Human Interactive Communication*. IEEE RO-MAN 2000 (Cat. No. 00TH8499), September 2000
9. Li, Z., Hauser, K.: Predicting object transfer position and timing in human-robot handover tasks. In: *Science and Systems* (2015)
10. Maeda, G.J., Neumann, G., Ewerton, M., Lioutikov, R., Kroemer, O., Peters, J.: Probabilistic movement primitives for coordination of multiple human-robot collaborative tasks. *Auton. Robots* **41**(3), 593–612 (2017)
11. Meyer zu Borgsen, S., Bernotat, J., Wachsmuth, S.: Hand in hand with robots: differences between experienced and naive users in human-robot handover scenarios. In: Kheddar, A., et al. (eds.) *Social Robotics*. ICSR 2017. LNCS, vol. 10652. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-70022-9_58

12. Nemlekar, H., Dutia, D., Li, Z.: Object transfer point estimation for fluent human-robot handovers. In: International Conference on Robotics and Automation (ICRA). Montreal, Canada (2019)
13. Perez-D'Arpino, C., Shah, J.A.: Fast target prediction of human reaching motion for cooperative human-robot manipulation tasks using time series classification. In: 2015 IEEE International Conference on Robotics and Automation (ICRA). IEEE, Seattle, WA, USA, May 2015
14. Shibata, S., Sahbi, B.M., Tanaka, K., Shimizu, A.: An analysis of the process of handing over an object and its application to robot motions. In: Computational Cybernetics and Simulation 1997 IEEE International Conference on Systems, Man, and Cybernetics, vol. 1, October 1997