Human Intention Recognition in Collaborative Environments using RGB-D Camera

Jelena Marić*,[1], Luka Petrović[1], Ivan Marković[1]

[1] University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
*maric298@gmail.com

***Abstract*—The increasing presence of robots and other autonomous systems in everyday environments, such as households, workplaces, schools and hospitals, requires their safe cohabitation and collaboration with humans. Accurate and reliable human intention recognition that fosters safe and efficient human-robot collaboration is thus a crucial component of these autonomous systems, especially if attainable from relatively cheap and widely available sensors such as RGB-D cameras. In this paper, we propose a hand-crafted model of human intention when reaching for one of multiple objects present on the table in front of a person. By coupling our hand-crafted model with a method for human skeleton tracking from RGB-D images, we devise a computationally efficient human intention recognition method suitable for collaborative pick-and-place scenarios. We experimentally verify our method in real-world scenarios of a person reaching for one of two, three and four objects placed on a table. We complement the paper with an open-source implementation of our human intention recognition method.**

***Keywords*—intention recognition, rgb-d, collaborative environments, action prediction, human-robot collaboration**

I. Introduction

As robots become more advanced and integrated into human-centric environments such as households, offices, schools and hospitals, there is a growing need for them to effectively collaborate and interact with humans, which presents new challenges in terms of system efficiency and human safety. Unlike robots, whose behavior can be fully controlled, human behavior is inherently unpredictable. Consider, for example, a person setting the dining table. While the task is well-defined, the order of subtask execution will clearly differ between people, especially if there are unforeseen changes in the environment. If the same task is carried out by a robot in collaboration with a person, it is of paramount importance for the robot to be aware of uncertainties and nuances of human behavior in order to ensure human safety and comfort. Human intention recognition thus presents one of the biggest challenges in collaborative robotics and has been an active area of research in recent years.

Intention recognition in the early stage of human actions can facilitate a variety of tasks in different scenarios, such as helping people overcome physical impediments [1], ensuring safety of human participants in traffic situations [2]–[4], enabling robots to purposefully partake in the manufacturing assembly process [5], or helping distinguish between healthy and unhealthy movement patterns in clinical settings [6]. The need for efficient human intention recognition is accentuated in collaborative environments where humans and robots simultaneously try to achieve a given task, especially in object-picking where a person and robot arm may often be in relative proximity. For such scenarios, a multimodal approach to intention recognition problem was presented in [7], where intention prediction is based on fusion of skeletal tracking measurements of a person's hand and pupil position measurements. Besides an RGB-D sensor used for obtaining skeletal data, additional piece of wearable equipment was needed for pupil tracking, reducing practicality of the proposed system.

Many researches have yielded successful results in intention recognition tasks using learning approaches [8], [9], where neural networks are often used as time series forecasters based on spatio-temporal movement patterns. Using kinematic data that includes motion and task-specific forces, a neural network was trained in [10] to approximate the function that represents dynamics of human movement and then predicts future human motions. Therein, intention refers to how the human hand will move next. A recurrent neural network (RNN) was trained in [11] to provide a prediction of future human movements given its history. Additionally, the RNN was then used for identifying the most probable goal of the predicted trajectory. In [12], the authors presented two methods for intention recognition that relied on RNNs, where the first method used person's hand trajectory as input, while the second method was trained on RGB-D videos. While both approaches showcased good performance when tested in an industrial setting, the high price and complex setup of motion tracking systems makes RGB-D sensor-based method more convenient and effective to use. A convolutional neural network (CNN) was proposed in [13] to recognize intention in each of the N captured images by learning subtle patterns in spatial features and accumulating those individual predictions over a time interval to produce the final prediction. Although spatial information in still images could be sufficient for action recognition, it might not be enough to distinguish subtle variations in motions which are a key aspect of intention recognition. To incorporate temporal aspect of motion, the authors in [14] proposed a CNN in a two-stream architecture which uses optical flow along with RGB data in an intention prediction task. A long short-term memory (LSTM) network combined with a CNN was used in [15] to predict reaching intention from gaze cues. In [16], [17], the

This research has been supported by the European Regional Development Fund under the grant KK.01.2.1.02.0119 (A-UNIT).

authors proposed using an ensemble of LSTMs for action prediction in collaborative environments, showcasing a degree of generalization to changes in the environment. While all of the mentioned methods achieve good performance in human intention recognition in collaborative environments, they require substantial amount of training data and hardware resources for the training process itself.

In this paper, we propose a lightweight human intention recognition method that uses RGB-D camera input. By relying on relatively cheap and widely available RGB-D sensors, as well as focusing on the ease of implementation, the proposed method is suitable for fast implementation and deployment on a variety of autonomous systems. Specifically, we craft a simple model of human intention when reaching for one of multiple objects present on the table in front of a person. Input data is processed in few simple steps using a limited amount of previously computed and stored values which makes the method computationally efficient. The proposed method considers distances between possible goal objects and key points of the tracked human skeleton, which are obtained from an existing method for RGB-D human skeleton tracking. The proposed method is suitable for collaborative pick-and-place tasks, and we experimentally verify its performance in real-world scenarios of a person reaching for one of two, three and four objects placed on a table. We complement the paper with an open-source implementation of our human intention recognition method[1].

II. Human intention recognition method

Throughout this paper, we consider a collaborative application where a human and a robot manipulate objects placed at known locations at the same table. In such scenarios, a person will typically move towards an intended goal, rather than in a random manner. Therefore, we need a human intention recognition method to determine a person's intended goal in order for the robot to choose its tasks appropriately, diminishing the risk of violating human safety and comfort. In this section we propose an easy-to-implement, computationally-efficient human intention recognition method that is suitable for the aforementioned scenario. We present each component of our method, including the definition of our weighted distance (Sec. II-A) and intention probability (Sec. II-B) functions, followed by the description of the utilized sliding window technique (Sec. II-C). Finally, our algorithm is summarized in Sec. II-D.

A. Weighted distance function

Consider a scenario where a person is grasping an object present on the table with their right hand. When a person begins their movement from a standing or sitting anatomical position, they start to reach with their hand towards the object. As the reaching motion continues, the distance between the target object and the person's hand becomes smaller and smaller. When the object is reached, that distance becomes zero. Moreover, the distance between target object and some of the other key points of the human skeleton, such as the right arm elbow or torso, becomes smaller with time.

We use this observation as the foundation of our intention recognition method. If there are multiple objects on the table, the distance between human skeleton keypoints and the preferred object at a certain time instant before grasping becomes smaller than the distance from every other object. Therefore, to be able to detect that the person is approaching the location of a certain goal g, we propose a general weighted distance function

$$d(g) = \sum_{j=1}^{J} w_j d_{gj} \tag{1}$$

where d_{gj} denotes the three-dimensional Euclidean distance between a goal location g and a given human skeleton joint j (out of J tracked skeleton joints), while w_j represents the weighting factor indicating relative importance of the given joint j, with $w_j \geq 0$, $\sum_{j=1}^{J} w_j = 1$.

While the proposed distance function is general in a sense that it can be used with any sensor that can provide skeleton detection, in this paper we employed a particular implementation of skeleton tracking from RGB-D sensor data. For data acquisition we used the widely available Xbox One Kinect sensor, that was initially intended for videogaming purposes, but quickly gained popularity in research applications due to its depth sensing ability and financial affordability. Since Kinect is a natural interaction device which can be accessed by open source APIs provided by OpenNI (Open Natural Interaction) framework, we used PrimeSense NiTE middleware which includes modules for OpenNI providing gesture recognition and skeleton tracking.

The mentioned skeleton tracking module provides position information for 15 different skeleton key points, namely hand, elbow and shoulder points for both arms, hip, knee and foot points for both legs, as well as head, neck and torso points. While this implies that $J = 15$, in our implementation of (1) we focus on the right hand, right elbow and torso information. We handcraft weights w_j in a way that right hand position information has the most influence on the total weighted distance. This corresponds to the aforementioned property of the distance between hand position and the desired object becoming zero when the object is grasped. The elbow and torso position information serve as a correction component in cases when the tracked human hand position provided by the used OpenNI framework is particularly inaccurate. We also assume that goal locations are invariable and known in advance, although they could be determined by using an open-source 3D object detection framework, e.g., [18].

B. Intention probability

When the weighted distance function in (1) is calculated for each possible goal location, we are in some cases able to simply infer the intended goal. For example, if we have three goals and one of them has comparatively much smaller weighted distance value than the other two possible goals, we may conclude that it is the intended

[1] https://github.com/maricjelena/intention-prediction

goal. Taking the argument of the minimum of the weighted distance function will provide reliable intention recognition in the aforementioned case that often happens when the person is already very close to the goal location. However, if the weighted distance is reasonably high from each possible goal, for example in the standing anatomical position, we might be unable to infer or be very uncertain about the intended goal. Taking the argument of the minimum in such situations may lead to clearly incorrect human intention estimates.

For the aforementioned reasons, we want to have a notion of uncertainty about our intention estimate and thus we utilize the *softmax* function to provide a probabilistic interpretation of the calculated weighted distances for each goal. A probability of each potential goal g_i (from a set of G possible goals) is given as

$$p(g = g_i) = \frac{e^{-\beta d(g_i)}}{\sum_{n=1}^{G} e^{-\beta d(g_n)}}, \tag{2}$$

where $d(g_i)$ is the weighted distance function given in (1), while β is a parameter for tuning the behavior of the softmax function, dependent on the specific requirements of a given task. With larger values of β, the function given in (2) has higher sensitivity. If we choose relatively large value of β, the probability of the predicted goal will always be close to 1, providing overconfident estimates. On the other hand, if chosen β is too small, it will output values close to 0.5 for the predicted goal, providing no confidence in the estimate. For the particular object reaching problem considered in this paper, we empirically concluded that $\beta = 40$ provides good estimates.

C. Sliding window

The only source of information we rely upon for intention recognition is skeleton tracking from RGB-D images, which is prone to providing noisy detections, e.g., due to self-occlusion. Unreliable input can significantly degrade estimation performance of the proposed intention estimation method. To mitigate the negative impact of wrong detections, we incorporate the sliding window technique in the estimation process to produce the final prediction formulated as the weighted moving average

$$\widetilde{p}_i = \frac{\sum_{f=1}^{F} w_f p_{i,f}}{\sum_{f=1}^{F} w_f} \tag{3}$$

where $p_{i,f}$ is the prediction probability of the goal g_i in the frame f, w_f is the weighting factor in the given frame f, and F is the number of frames considered in the weighting group. Parameters w_f and F (i.e., sliding window size) are manually tuned to work well for the observed motion.

Generally, applying the moving average with time-series data smooths out short-term fluctuations. In the context of this work, noticeable and sudden changes in tracked position do not occur naturally, because it is not physically possible for people to significantly change position of any joint in short period of time such as 0.03 s. However, this kind of deviation is not rare in skeleton detection. The purpose of the sliding window technique is to smooth out detection errors by predicting the intention using weighted contributions of the last F predictions, including the current one. Earlier predictions contribute less than the newer ones to better represent the time connection of the tracked position changes during the motion. For the particular problem considered in this paper, we empirically chose $F = 7$ which provided good balance between stability and fast prediction times in our experiments.

D. Algorithm summary

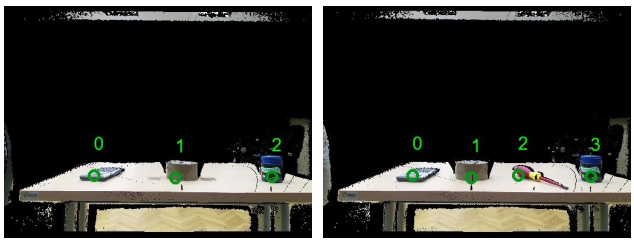
Given the skeletal pose of a human in a frame f, we first compute the weighted distance function given in (1) for each goal location g_i based on the current measurement of the considered joint positions. We then calculate the probabilities for each goal g_i following (2). Afterwards, the sliding window is recursively applied using (3) to get the final probability distribution at the current time instant. If there are not enough previous frames when estimating the intention in the current frame f, i.e., $f < F$, then the predictor takes f frames into account with the adjusted weight of each frame. The object location with the highest associated probability is considered to be the estimated intended goal.

III. Experimental results

In this section we demonstrate the performance of our method on data we collected in our laboratory. We start by explaining our dataset collection process and the obtained data in Sec. III-A. Then we present the results of the conducted performance analysis for scenarios with differing levels of prediction difficulties determined by the number of objects and their spatial arrangement. We first utilize the proposed intention prediction method in a two-object scenario (Sec. III-B) to verify the applicability of the proposed method in a fairly simple intention discrimination task. Experimentation area is then expanded by adding more objects (Sec. III-C), since the intention prediction method can be of great use in a real-life application based on recognizing the intended goal among many others. This also gives us an opportunity to induce possible problems that rarely occur in the basic two-object scenario and to show how they are handled with the proposed method. Lastly, in Sec. III-D we observe the effect of the sliding window technique on the stability of the proposed method when detection errors occur.

A. Dataset collection

We recorded 3D coordinates of a person's skeleton position while the person was reaching for objects placed on a table at least 9 cm apart. The number of those objects $K \in \{2, 3, 4\}$ and their locations are known for each scenario in advance and are depicted in Fig. 1. Each recorded sequence includes data acquired during multiple reaching movements. The duration of each reaching motion is in range of $1.5 - 2$ s, starting from the neutral body posture and ending in object reaching, after which the person returns to the neutral position or continues to reach for an another object directly from the location of the previously reached object. Sequences are labeled with timestamps of moments when an object is reached.



(a) (b)

Fig. 1: Scene with: (a) 3 and (b) 4 objects

B. Toy problem: Two object scenario

The method was first tested in the basic case of two objects being located with maximum distance apart (Fig. 2). As shown in Fig. 3, intention is correctly predicted for each reaching motion with a high probability of the intended object throughout the whole sequence. This is a somewhat simple decision because of the arrangement of objects. They are positioned such that approaching one object increases the distance to the other one which means that individual direction of probability change can point directly to the estimate of the method.

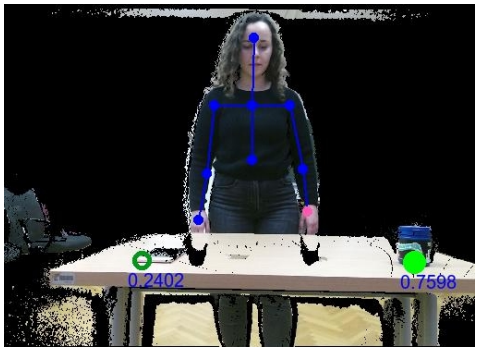


Fig. 2: Scene with two objects located with a maximum distance

C. Multi-object scenario

In this section we examine the prediction performance in scenarios featuring two, three and four possible goal objects. A comparison of prediction performance for three different combinations of tracked human skeleton joints being used in probability computation is given in Table I. For the right hand and right elbow combination, we used $w_j = 0.8$ for the hand and $w_j = 0.2$ for the elbow. For the right hand, right elbow and torso combination, we used $w_j = 0.8$ for the hand, and $w_j = 0.1$ for both the elbow and the torso. The table of prediction results shows average time left between the time instant of intention recognition and a manually determined time instant of reaching the goal location. This time is calculated by averaging results for chosen reaching motions in the observed sequence. Sequence selection criteria are explained in the following paragraph.

A given time instant is considered to be the prediction time instant if: i) the probability $\widetilde{p}_i$ of the goal g_i is greater than 0.5 in scenarios containing 2 objects and greater than 0.4 in scenarios containing 3 or 4 objects

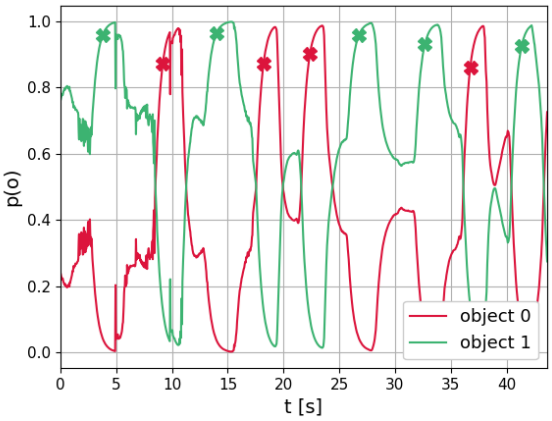


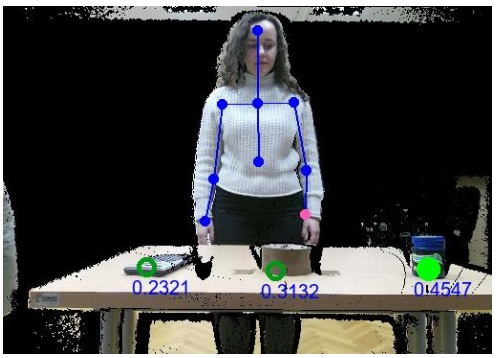
Fig. 3: Probabilities of objects from sequence shown Fig. 2 being reached. Note: moments when the goal objects are reached are marked with 'x' coloured with the ground truth object's color.

TABLE I: Average time left between the moment when intention is recognised and the reaching moment. Note: higher value represents better performance.

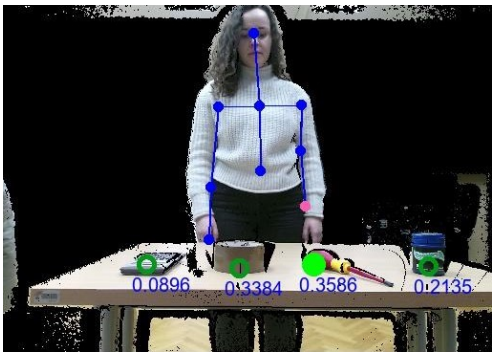
Reference skeleton joints	Number of objects		
	2	3	4
right hand	0.7559	0.6472	0.4789
right hand + right elbow	0.7219	0.6029	0.4152
right hand + right elbow + torso	0.7570	0.6279	0.4727

(Fig. 4); and ii) probability $\widetilde{p}_i$ of the goal g_i being reached is constantly greater than other objects' probabilities until (at least) the moment of reaching. There are cases when the initial position of a person suggests that a certain object is being reached, i.e., the initial probability of an object is higher than the defined threshold (Fig. 4a). Reaching that particular object is ignored in the analysis because the results do not reflect methods true prediction ability. The unevenness of initial probabilities could be alleviated with predefined location of a person such that the body position is truly neutral. Reaching motions which continue from the previous goal location were also not taken into account here due to significant differences in duration in comparison to trajectories starting from a neutral position which made them unsuitable for direct comparison. Note that the probability threshold of 0.4 was empirically determined as a suitable value for scenarios with three and four objects due to task complexity increase with the increase in the number of objects. An example of introducing complexity in an estimation task is a situation when approaching intended object results in comparable distance decrease for other objects as well.

The results from Table I show that the proposed method can provide reliable human intention recognition up to $0.75\,\mathrm{s}$ before reaching the desired object. Expectedly, the increased number of objects leads to decreased performance in average time left to correct intention recognition. These results also indicate that different joint combinations lead to similar results in the general case, implying that the intention could be determined by tracking only the hand joint, which is true for the chosen sequences. However, it is clear that the prediction performance declines when the detection does not correspond with the true position



(a)



(b)

Fig. 4: Scenarios containing (a) 3 and (b) 4 goal locations marked with initial probabilities

TABLE II: Average time left between the moment intention is recognised and the reaching moment in a scenario with 3 objects where input data suffers from detection errors.

Reference skeleton joints	Δt
right hand	0.5830
right hand + right elbow	0.5327
right hand + right elbow + torso	0.5940

of a person, as is often the case for hand joint. Taking the elbow and torso measurements into account can enhance output stability when hand detection significantly changes due to unreliable tracking or possible occlusions. To support this claim, we chose another sequence, depicted in Fig. 5, to show how the suggested sets of joints affect the final prediction. The average time left before correct intention recognition for the pertaining sequence is shown in Table II and suggests that the utilization of different tracked body joints may improve performance when input data suffers from detection errors.

D. Sliding window impact

We conducted additional performance analysis to show the importance of using sliding window. For that purpose, a sequence containing detection errors in multiple frames within one sliding window was chosen. Our analysis showed that, when using the sliding window, intention prediction method has the ability to resist negative influence of few false detection inputs, as showcased in Table III. The visual representation of prediction changes before and after the application of the sliding window technique is shown in Fig. 6. If skeleton detection was consistently

TABLE III: Probabilities when reaching for object $i = 1$ in 8 consecutive frames from Fig. 5 depending on the use of sliding window. Number of possible goal locations is 3, i.e., $i \in \{0, 1, 2\}$ seen from left to right. Predictions which differ from the ground truth are marked red.

time instant t	no sliding window		sliding window	
	i	$max(p_i)$	i	$max(\widetilde{p}_i)$
t_k	1	0.8431	1	0.5184
t_{k+1}	1	0.9188	1	0.5528
t_{k+2}	0	0.7632	1	0.4371
t_{k+3}	1	0.9508	1	0.5752
t_{k+4}	1	0.9526	1	0.5999
t_{k+5}	1	0.7210	1	0.5734
t_{k+6}	0	0.8137	1	0.4790
t_{k+7}	0	0.7510	1	0.4763

bad, then the prediction would have also underperformed since we rely solely on these measurements. The overall performance of the proposed method depends greatly on the provided input which can be unreliable since a single Kinect is used for obtaining the data. Introducing one or more additional RGB-D cameras would offer better scene coverage, hence improving the fused measurement accuracy.

IV. Conclusion

In this paper, we proposed a simple skeletal data-based model for recognizing intention in human actions when reaching for one of multiple objects present on the table in front of a person. By coupling our hand-crafted model with a method for human skeleton tracking from RGB-D images, we devised an intention recognition method suitable for collaborative pick-and-place scenarios. We experimentally verified the performance of the proposed method in real-world scenarios of a person reaching for one of two, three and four objects placed on a table. The proposed method can also be adapted for usage in different scenarios by picking a subset of skeleton joints with weights adjusted accordingly to the nature of the task. The obtained results support the decision of considering multiple joints in the distance function along with the sliding window technique for smoothing the prediction curve, both ensuring the stability of prediction method to some extent. If objects were placed in a straight line between the human and the farthest object, ambiguity of movement path would lead to method becoming greedy and declaring the location of the nearest object as the intended one; therefore, the proposed method is not suitable for every arrangement of objects.

In future work, it would be interesting to implement a robot controller for a human-robot collaborative assembly task to evaluate the performance of presented method in an real-life industrial environment.

References

[1] C. A. Chin, A. B. Barreto, J. G. Cremades, and M. Adjouadi, "Integrated electromyogram and eye-gaze tracking cursor control system for computer users with motor disabilities." *Journal of rehabilitation research and development*, vol. 45 1, pp. 161–74, 2008.

[2] H. Berndt, J. Emmert, and K. Dietmayer, "Continuous driver intention recognition with hidden markov models," in *2008 11th International IEEE Conference on Intelligent Transportation Systems*. IEEE, 2008, pp. 1189–1194.

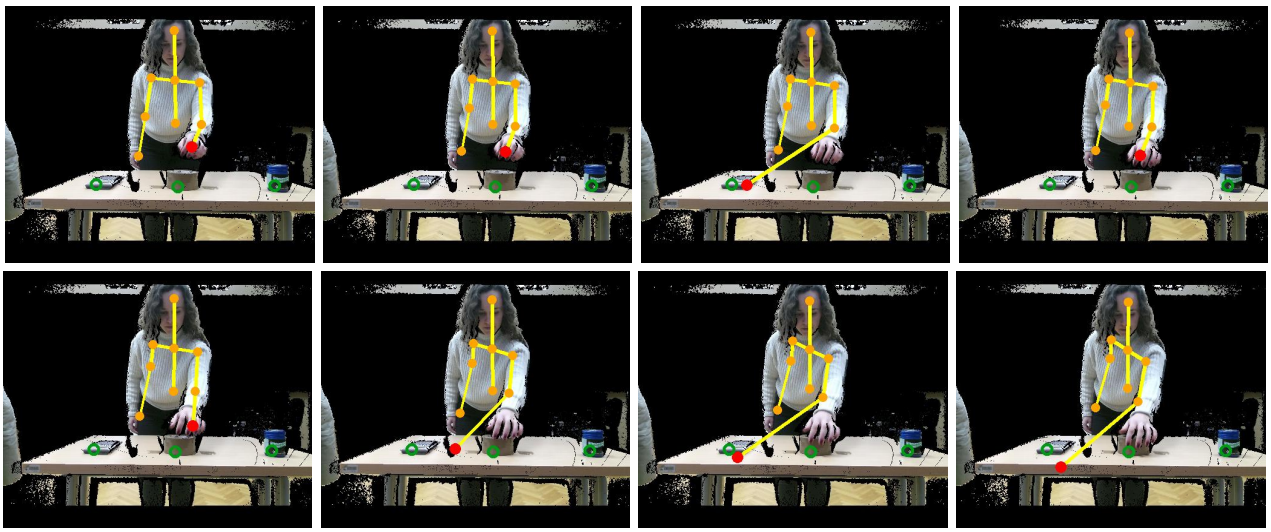
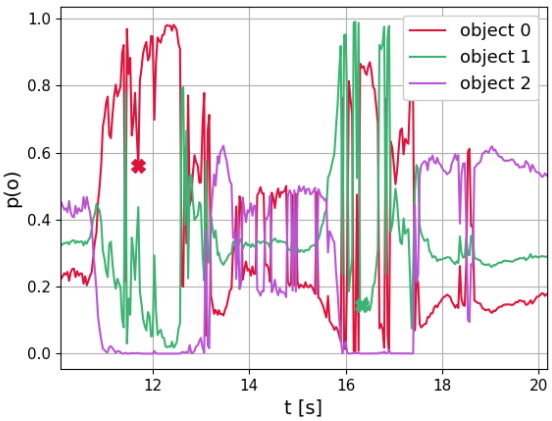
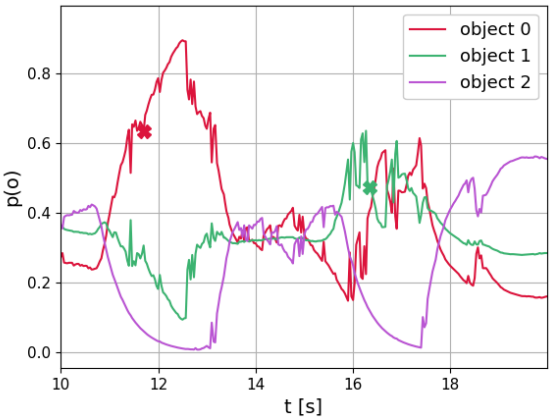


Fig. 5: Skeleton detection in consecutive frames



(a)



(b)

Fig. 6: Probabilities of reaching objects (a) without and (b) with sliding window.

[3] Z. Fang and A. López, "Intention recognition of pedestrians and cyclists by 2d pose estimation," *IEEE Transactions on Intelligent Transportation Systems*, vol. PP, pp. 1–11, 10 2019.

[4] F. Camara, N. Bellotto, S. Cosar, F. Weber, D. Nathanael, M. Althoff, J. Wu, J. Ruenz, A. Dietrich, G. Markkula *et al.*, "Pedestrian models for autonomous driving part ii: high-level models of human behavior," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 9, pp. 5453–5472, 2020.

[5] Z. Zhang, G. Peng, W. Wang, Y. Chen, Y. Jia, and S. Liu, "Prediction-based human-robot collaboration in assembly tasks using a learning from demonstration model," *Sensors*, vol. 22, no. 11, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/11/4279

[6] F. Ragni, L. Archetti, A. Roby-Brami, C. Amici, and L. Saint-Bauzel, "Intention prediction and human health condition detection in reaching tasks with machine learning techniques," *Sensors*, vol. 21, no. 16, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/16/5253

[7] D. Trombetta, G. S. Rotithor, I. Salehi, and A. P. Dani, "Human intention estimation using fusion of pupil and hand motion," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 9535–9540, 2020, 21st IFAC World Congress. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S240589632033113X

[8] P. Schydlo, M. Rakovic, L. Jamone, and J. Santos-Victor, "Anticipation in human-robot cooperation: A recurrent neural network approach for multiple action sequences prediction," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 5909–5914.

[9] P. Kratzer, N. B. Midlagajni, M. Toussaint, and J. Mainprice, "Anticipating human intention for full-body motion prediction in object grasping and placing tasks," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2020, pp. 1157–1163.

[10] E. C. Townsend, E. A. Mielke, D. Wingate, and M. D. Killpack, "Estimating human intent for physical human-robot co-manipulation," *arXiv preprint arXiv:1705.10851*, 2017.

[11] D. Nicolis, A. M. Zanchettin, and P. Rocco, "Human intention estimation based on neural networks for enhanced collaboration with robots," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 1326–1333.

[12] M. Mavsar, M. Deniša, B. Nemec, and A. Ude, "Intention recognition with recurrent neural networks for dynamic human-robot collaboration," in *2021 20th International Conference on Advanced Robotics (ICAR)*. IEEE, 2021, pp. 208–215.

[13] L. Zhang, S. Li, X. Diao, and O. Ma, "An application of convolutional neural networks on human intention prediction," *International Journal of Artificial Intelligence Applications*, vol. 10, pp. 1–11, 09 2019.

[14] S. Li, L. Zhang, and X. Diao, "Deep-learning-based human intention prediction using rgb images and optical flow," *Journal of Intelligent & Robotic Systems*, vol. 97, pp. 95–107, 2020.

[15] P. Festor, A. Shafti, A. Harston, M. Li, P. Orlov, and A. A. Faisal, "Midas: Deep learning human action intention prediction from natural eye movement patterns," *ArXiv*, vol. abs/2201.09135, 2022.

[16] T. Petković, L. Petrović, I. Marković, and I. Petrović, "Ensemble of lstms and feature selection for human action prediction," in *Intelligent Autonomous Systems 16: Proceedings of the 16th International Conference IAS-16*. Springer, 2022, pp. 429–441.

[17] ——, "Human action prediction in collaborative environments based on shared-weight lstms with feature dimensionality reduction," *Applied Soft Computing*, vol. 126, p. 109245, 2022.

[18] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3d object detection and tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 784–11 793.