# Context attention: Human Motion Prediction using context information and deep learning attention models

Javier Laplaza, Francesc Moreno-Noguer, Alberto Sanfeliu***

Universitat Politècnica de Catalunya, Catalonia, Spain
javier.laplaza@upc.edu

**Abstract.** This work proposes a human motion prediction model for handover operations. The model uses a multi-headed attention architecture to process the human skeleton data together with contextual data from the operation. This contextual data consists on the position of the robot's End Effector (REE). The model input is a sequence of 5 seconds skeleton position and it outputs the predicted 2.5 future seconds position. We provide results of the human upper body and the human right hand or Human End Effector (HEE).
The attention deep learning based model has been trained and evaluated with a dataset created using human volunteers and an anthropomorphic robot, simulating handover operations where the robot is the *giver* and the human the *receiver*. For each operation, the human skeleton is obtained using OpenPose with an Intel RealSense D435i camera set inside the robot's head. The results show a great improvement of the human's right hand prediction and 3D body compared with other methods.

**Keywords:** machine learning, human-robot collaboration

## 1 Introduction

In order to further integrate robots in human society, they are required to better understand how humans move and interact with the environment. Both Robot-Human interaction [7] and Robot-Human Collaboration [19] require a set of skills, such as perception, navigation or anticipation. In this work we will focus on this last skill.

Humans anticipate the movement of other humans when they need to interact with them. Delivering a tool, practicing sports, playing games or simply opening a door for a someone else are some examples of activities that involve human motion prediction. Our goal is to improve the quality of human-robot interaction by enabling robots to predict human motion.

* All authors work in the Institut de Robòtica i Informàtica Industrial de Barcelona (IRI), Catalonia, Spain jlaplaza, fmoreno, sanfeliu@iri.upc.edu

** Work supported under the Spanish State Research Agency through the ROCO-TRANSP project (PID2019-106702RB-C21 / AEI / 10.13039/501100011033)) and the EU project CANOPIES (H2020- ICT-2020-2-101016906)

For this reason, we focus on prediction related to a human-robot handover task (see Fig. 1). We aim to fuse information from the human skeleton and contextual data from the operation (Sec. 3.2). Regarding the human skeleton we fuse data from the whole upper body skeleton and data from each part of the upper body separately (left arm, right arm and middle body). Regarding context we fuse data from the Robot End Effector (REE).

Our approach allows us to add as much information queues as desired since we use a multi-head attention channel to focus independently on each one of these queues, so further context information could be added, such as the human gaze or the position of obstacles placed in the scenario.

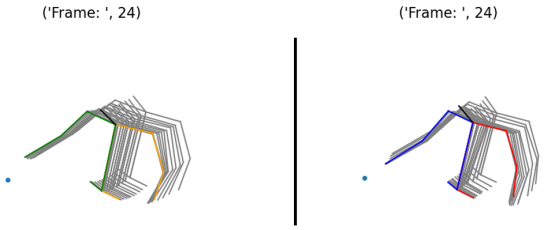


**Fig. 1.** Left: Model predicted trajectory. Right: Ground truth trajectory. In both cases, the blue dot shows the REE location and the motion trajectory in gray. Both skeletons start from the same position.

The data used to train the model was collected in our own laboratory, using an anthropomorphic robot named IVO (see Fig. 2) and a group of human volunteers. 7 human volunteers participated in the data collection. Each one of these volunteers recorded several sequences of a handover operation, featuring different behaviors during the operation.

In section II we explain the related work. In section III the model architecture is detailed, for both the skeleton and context information. The dataset creation is described in section IV. In section V we discuss the experimental results and finally, in section VI we draw some conclusions.

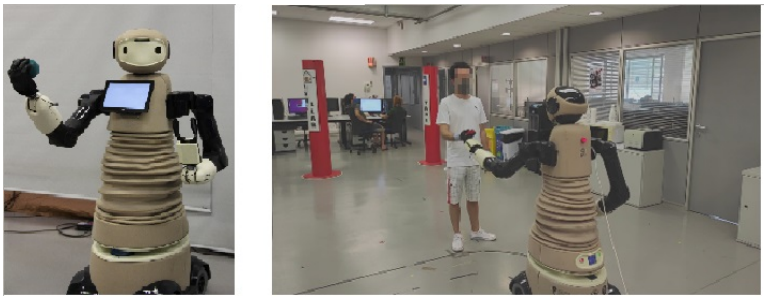

**Fig. 2.** The IVO robot shown in the figure has been used to create a human-robot handover dataset used to train our model.

## 2 Related Work

There are interesting attempts to include contextual information in human motion prediction.

The approach from [6] is philosophically very similar, since the model predictions are conditioned on the objects around the humans, such as tables or doors. The model uses a GAN architecture to exploit this added information.

Another very remarkable work is the one presented by [17], where they use Transformer VAE, which also uses attention, to predict the human motion, but they condition their prediction with the action that the human is performing, which can also be considered as context.

If we look at the human motion prediction field in a wider sense, we can find different approaches that take advantage of different model architectures.

In [15] by Martinez et al., the problem is approached as a time series algorithm, proposing a RNN architecture able to generate a predicted human motion sequence given a real 3D joint input sequence. Although the results obtained in this model are quite promising, the work raises attention in very particular case: a non-moving skeleton can often improve results in a L2 based metric. This is commonly the most studied approach, used in [8], [11], [1].

The work done in [4] by Bütepage et al. shows how advances in latent variable models such as Variational Autoencoders can be used in order to produce relevant results. In this work, the upper body motion is predicted up to 1660 ms. The main idea is to predict the future time steps given some previous time steps. Thus, a joint probability is modeled, using these two variables and a number of hidden variables who governs the unobserved dynamics.

In [2], Barsoum et al. take a similar approach, modifying the structure to introduce GANs. By feeding the network with a skeleton input sequence plus a random z vector drawn from a uniform or Gaussian distribution $z \sim p_z$, a predicted sequence is computed. They add two losses to the architecture to try to get consistent skeletons in their predictions: consistency loss (to ensure that no drastic movements between frames appear) and bone loss (to ensure that

bone lengths from the predicted skeleton don't change). In [9], GANs are used to reconstruct skeletons in sequences with occlusion problems.

The same study of the field can be done in the robotics side, where there have been previous attempts to introduce the prediction in human-robot tasks, specifically handovers.

In [10], Hoffman et al. compare anticipatory versus reactive agents. The first methods tend to feel more fluent and natural to humans that collaborate with robots, stressing the importance of being able to predict the intention of the human partner.

In [12], Lang et al. use a Gaussian Process clustered with a stochastic classification technique for trajectory prediction using an object handover scenario. Real 6D hand movements are captured during human-human handovers to classify the grasping position of both humans using a maximum likelihood estimation. Although obtaining interesting results, our goal is to obtain the motion of all the body joints or, at least, the upper body joints. Furthermore, we argue that using human-human datasets would not represent the real behavior of a human moving around a robot. Other studies about the handover task which focus on human-human handovers are [16] and [3].

## 3 Model Architecture

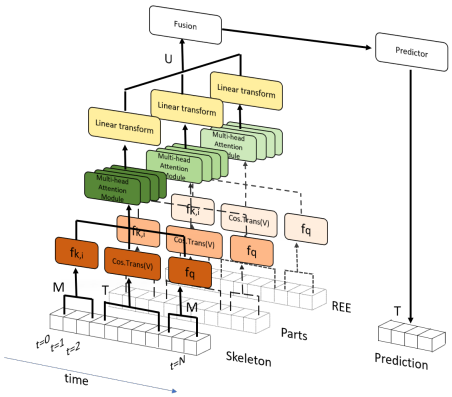


**Fig. 3.** We modify the original model to add additional information using C channels. Each channel experiments the same operations. First the data from each channel are embedded using the function $f_{k,i}$, $f_q$ and the Discrete Cosine Transform (orange layer). The output of each funcion is fed to the corresponding multi-head attention module (green layer), which is then passed through a linear transformation (yellow layer) before being fused with the outputs of the other information channels. Then, the fused output is fed to the predictor, where the predicted skeleton trajectory is generated.

The model proposed here (Fig. 3) is inspired in the work published by Mao et al. [14] where we have introduced some modifications to achieve our goals. We change the attention module by a multi-head attention module, in a similar way that is used in the original Attention paper by Vaswani et al. [18]. By doing this, we increase the attention capacity of the model, allowing the network to exploit more patterns from the data. We also modify the architecture to add multiple attention channels side by side that allow the model to use data from different sources. Finally, the processed information is then merged following different strategies and then fed into a predictor module to obtain the human skeleton prediction.

The model uses an attention deep learning based neural network able to discover sub-sequences inside the main sequence.

Let us consider $X_{1:N} = [x_1, x_2, x_3, ..., x_N]$ as the input sequence consisting of N human upper body skeleton 3D poses $x_i \in \mathbb{R}^k$ for each time frame $i$, where $K$ is the number of parameters required to represent the human pose. The model goal is to predict the future T 3D upper body skeleton poses, $X_{N+1:N+T}$.

We also consider a vector $X_P = [x_{P,1}, x_{P,2}, x_{P,3}, ..., x_{P,N}]$ $X_P \in \mathbb{R}^{3,k}$, where data from each body part are considered independently. The body parts defined were the right arm (right wrist, right elbow and right shoulder), the left arm (left wrist, left elbow and left shoulder) and the middle body (right hip, left hip, chest and head).

Similarly, a vector $X_E = [x_{E,1}, x_{E,2}, x_{E,3}, ..., x_{E,N}]$ $x_{E,i} \in \mathbb{R}^3$ defines the position of the REE during the same time frames.

Since the model goal is to predict the future T poses given the M previous poses, the data can be arranged to comply with the classical attention formulation: all the recordings are divided into sub-sequences of $M + T$ frames, creating $N - M - T + 1$ sub-sequences $\{X_{i:i+M+T-1}\}_{i=1}^{N-M_T+1}$.

Each of these sub-sequences consist on a *key* (the first M) and a *value* (the whole M+T sub-sequence), composing a key-value pair. The last M frames from the input sequence is considered to be the *query* and will be used to predict the following T frames the same way the network was trained to do in all the previous sub-sequences.

### 3.1 Multi-headed Attention

We define a channel $C_i$ with a multi-head attention module for each input sequence $X, X_P$ and $X_E$, each channel consisting on $N_{heads}$ heads. Each sequence is fed to its corresponding multi-head attention module, and each head of the multi-head performs the classical scaled dot product operation, shown in Eq. 1 to compute the attention scores ($a$):

$$a_{C,N_{heads},i} = \frac{qk_i^T}{\sum_{i=1}^{N-M-T+1} qk_i^T} \tag{1}$$

Before this operation, the query ($q$) and keys ($k$) are mapped to vectors of the same dimension $d$ with two functions $f_q : \mathbb{R}^{KxM} \rightarrow \mathbb{R}^d$ and $f_k : \mathbb{R}^{KxM} \rightarrow \mathbb{R}^d$, modeled with convolutional neural networks:

$$q = f_q(X_{N-M+1:N}), k_i = f_k(X_{i:i+M-1}) \tag{2}$$

Where $q, k_i \in \mathbb{R}^d$ with $i \in \{1, 2, ..., N - M - T + 1\}$.

The model also maps the values ($V$) to trajectory space using a Discrete Cosine Transform on the temporal dimension.

The output of the attention model ($U$) is then computed as the weighted sum of values:

$$U_{C,N_{heads}} = \sum_{i=1}^{N-M-T+1} a_{C,N_{heads},i} V_{C,N_{heads},i} \tag{3}$$

Where $U \in \mathbb{R}^{k,(M+T)}$.

Then, the output of each head is fed into a single linear transformation layer $h$:

$$U_C = h(U_{C,1} \parallel U_{C,2}, ... \parallel U_{C,N_{heads}}) \tag{4}$$

Where $\parallel$ signals the concatenation of the outputs of each head.

### 3.2 Information Channels Fusion

Each information channel $C_i$ computes its own attention scores related to the data it was fed (whether skeleton or context) and then the outputs are fused.

We took three different strategies to fuse the data:

1. Directly concatenating the scores $U = U_1 \parallel U_2 ... \parallel U_C$ and then feeding the concatenated outputs to a predictor module in charge of the prediction.
2. Each channel output is weighted by a trainable parameter $W = W_1, W_2, ..., W_C$ and merged with the rest. The weighted output is then passed to the predictor module in order to estimate the future skeleton poses.
3. Each channel has a corresponding predictor module, which outputs a different trajectory. These trajectories are then weighted and merged to obtain a result trajectory.

The fusion strategy that yielded better results was the second one, so we opted to use it for our experiments.

### 3.3 Predictor Module

We used the same predictor module than Mao et al. [14]. The predictor module uses the discrete cosine transform representation to encode temporal information of each joint coordinate and graph convolutional networks with learnable adjacency matrices to learn the spatial dependencies among them.

Thus, the predictor output is the last M frames of the input sequence followed by the predicted T frames encoded in the frequency domain. By using an Inverse Discrete Cosine Transform (IDCT) we obtain the skeleton poses in cartesian coordinates.

### 3.4 Loss

We use a $L^2$ loss function to minimize the cartesian distance between our predictions and the ground truth data:

$$\mathcal{L}(\hat{p_{t,j}}, p_{t,j}) = \frac{1}{J(M+T)} \sum_{t=1}^{M+T} \sum_{j=1}^{J} \|\hat{p_{t,j}} - p_{t,j}\|^2 \tag{5}$$

Where $p_{t,j}$ is the ground truth position and $\hat{p_{t,j}}$ is the output trajectory from the predictor.

## 4 Dataset

In our previous work [13] we created a custom dataset in our laboratory. Here we used the same dataset in order to compare the new model to the previous one.

The dataset was collected using the anthropomorphic robot IVO and human volunteers performing a handover task where the human is the *receiver* and the robot the *giver* (see Fig. 4). The human and the robot approach towards each other and extend their arms to reach their partner. The delivered object is a 10 cm sided cube handed to the human using the robot left arm. The human always picks the object using the right arm.

A video of each sequence is recorded using an Intel RealSense D534i camera placed inside the robot's head. The video is recorded at a framerate of 10 fps. The recording is finished when the human is about to remove the object from the robot end effector REE.

**Fig. 4.** Example of two sequences recorded for the dataset. Left: Third person view of the experiment. Right: Robot point of view, showing the detected skeleton in the image.

The skeleton of the human is obtained from each sequence using OpenPose (Cao et al. [5]) to extract the 2D joint locations on the image. These 2D joints and the camera depth map data are used to obtain the 3D coordinates of each joint.

Only the upper body (from the hips to the head) of the human is used to avoid occlusions of the legs when the human is close to the robot.

The volunteer recreates five different behaviors: (1) picking the object standing close to the robot from the beginning (*close*); (2) picking the object as they would naturally do (*natural*); (3) picking the object delaying the arm motion once they are in range to pick the object (*delay*); (4) picking the object and then holding the hand still with the object grabbed (*hold*); and finally, (5) picking the object doing a free arm movement, while he/she approaches, such as checking their smartphone, waving their hands or stretching.

The robot also performed three different behaviors: the robot could be offering the object from the beginning, the robot could offer the object while the human was approaching, or the robot could approach to the human while simultaneously offering the object while the human was approaching.

Once all the sequences were recorded, we performed a sanity check of the data by visual inspection.

We used seven volunteers (3 women and 4 men, ages ranging from 25 to 60 years old) to perform the recordings. Each volunteer records all the possible scenarios, 15 scenarios in total, repeating each scenario once, which means 30 sequences for each volunteer, 210 sequences in total, ranging from 4 to 30 seconds. Considering that we use sub-sequences of 75 frames in the model and that data was visually inspected to discard corrupted data, we end up with 7.214 samples using data augmentation, each one containing 75 frames.

Depending on the scenario, the human initial position is 1.3 meter in front of the robot (close scenarios), 5 meters (scenarios where the robot moves towards the human) or 3 meters(the rest of cases), with no obstacles between the robot and the human.

## 5 Experimental Results

### 5.1 Experimental details

We use our dataset and split the subjects in training dataset (subjects 2 to 7) and validation dataset (subject 1).

For training, we use 50 frames (5 seconds) as input and output 25 frames (2.5 seconds). We choose this time windows to compare directly with the model presented in [14]. We perform an ablation study considering each single feature of the model separately, more specifically how the number of heads and the channels affect the results.

In order to compare with other methods, we train and validate other human motion prediction models in our dataset. All the results shown in Table 1 are obtained using our training and validation dataset.

### 5.2 Experiments

We compute the $L_2$ distance in Cartesian coordinates between our predicted sequences and the ground truth sequences for the same input sequence. Table 1 contains the computed errors along the test dataset before overfitting over the training dataset.

We also compute how many frames in the sequence have an error equal or less than 0.15m and 0.25m, and give the percentage of successful frames.

Finally, we check the $L_2$ error for the right hand of the human (HEE), since it is the most important joint in the handover task.

The first two rows of the table corresponding to the (*RNN*) and the (*Hist. Rep.Itself*) are used as baseline. We use the models presented in [15] and [14] respectively and train them in our dataset. Note that the disparity in the results shown here with the presented in their corresponding papers comes from the differences between our dataset and the datasets where these models were trained (the H36M dataset), where actions and framerates are quite different.

From third to seventh row we show the results presented in our previous work [13].

From the eighth row to the last one, we show the results from our ablation study. We decided to compare the multi-head attention module using 1 head and 4 heads, since it is a way to compare the model with and without the "multi-head" feature (the case with 1 head).

Some remarks should be made before discussing the results: the row showing the better accuracies is the seventh row, corresponding to a configuration evaluated in [13]. This configuration evaluated a certain model using only data from the skeleton that was really close to the robot, so the motion of those sequences was relatively small. In all the configurations evaluated in this work we use the entire sequence, hence our results are slightly worse in the shown metrics.

One remarkable feature from the results is that the average accuracy of all the joints has increased significantly compared to the other models, obtaining accuracies ranging from 16.8 to 13,4 cm compared with the accuracies around 20 cm obtained by the other models. We believe that this improvement is related to a smarter use of the attention structure used in the presented model. Scores obtained in the number of samples with errors below 15 and 25 cm have also been improved, and more importantly, the right hand accuracy has also improved.

In our ablation study we see that using the multi-head module usually increases the accuracy of the results. Part conditioning doesn't seem to improve significantly the accuracy of the model, although it scores the best in the "error below 25 cm" metric.

The addition that seems to obtain better scores overall is the one that adds only the information of the robot end effector, obtaining and error of 13.5 cm in the overall joint metric and a 17.7 cm error for the right hand metric.

We also did some preliminary implementation of the model in the IVO robot in order to check if the model returns meaningful results. Thus, the model was encapsulated in a ROS node and tested with 2 human volunteers that didn't participate in the dataset collection.

The results obtained in this real experiment showed realistic predictions while the human was approaching the robot, allowing the robot to use this predicted motion (Fig. 5).

| **Model** | $L_2$ (m) | % $\leq$ 0.15m | % $\leq$0.25m | Right Hand $L_2$ (m) |
|---|---|---|---|---|
| RNN [15] | 1.19 | 4.35 | 12.78 | 1.45 |
| Hist. Rep. Itself [14] | 0.213 | 56.03 | 70.82 | 0.348 |
| End Effector conditioning [13] | 0.207 | 58.67 | 72.78 | 0.349 |
| Prob. Distr. modelling [13] | 0.224 | 58.78 | 71.21 | 0.365 |
| End Effector cond + Prob. Distr. modelling [13] | 0.221 | 68.16 | 76.97 | 0.264 |
| End Effector cond + Prob. Distr. modelling (Approaching) [13] | 0.222 | 66.35 | 76.18 | 0.228 |
| End Effector cond + Prob. Distr. modelling (Pre-contact) [13] | 0.100 | **85.61** | 91.5 | 0.073 |
| Multi-head module (1 head) | 0.168 | 72.37 | 83.72 | 0.219 |
| Multi-head module (4 heads) | 0.152 | 73.07 | 85.09 | 0.206 |
| 1 head + Body Part channel | 0.151 | 71.98 | **87.93** | 0.210 |
| 4 head + Body Part channel | 0.155 | 73.58 | 84.71 | 0.211 |
| 1 head + End Effector Channel | 0.135 | **73.93** | 84.67 | **0.177** |
| 4 head + End Effector Channel | 0.139 | 73.22 | 83.77 | 0.189 |
| 1 head + Body Part channel + End Effector Channel | 0.137 | 72.96 | 83.73 | 0.186 |
| 4 head + Body Part channel + End Effector Channel | **0.134** | 73.18 | 84.28 | 0.186 |

**Table 1.** Results obtained across the validation dataset.

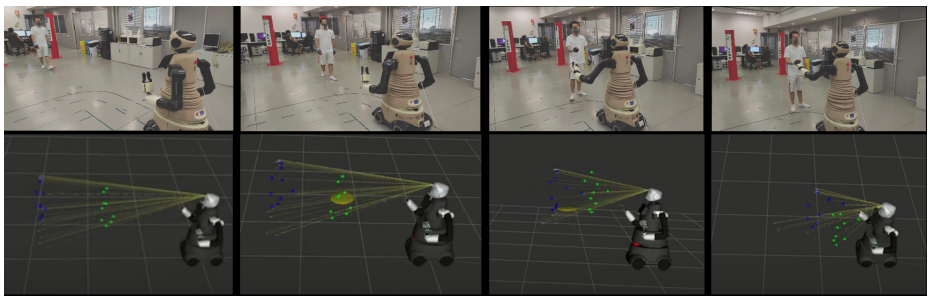

**Fig. 5.** We tested the predictor on the real robot during a handover operation. Top: Video sequence of the operation. Bottom: Visualization with ROS of the predictor output (Blue dots are the current human position, green dots are the predicted human position.

## 6 Conclusions

We presented an attention based neural model to characterize the motion of a human skeleton 2.5 seconds in the future, performing a handover task with a robotic partner and obtaining the future human motion predictions using contextual information, specifically the human body parts and the position information of the REE.

We proposed a modular approach to add contextual queues to the model to enhance predictions in handover tasks, but the same idea can be extrapolated to other tasks and new contextual information such as gaze or obstacle positions.

We obtained better results than previous models both for the average body joints and the human right hand.

Futhermore, we implemented the prediction model in the IVO robot and obtained feasible predictions replicating handovers with a small group of humans. This opens a future research line where the convenience of using the prediction of the human during collaborative tasks can be explored.

## References

[1] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. "Structured Prediction Helps 3D Human Motion Modelling". In: *CoRR* abs/1910.09070 (2019). arXiv: 1910.09070. URL: http://arxiv.org/abs/1910.09070.

[2] Emad Barsoum, John Kender, and Zicheng Liu. "HP-GAN: Probabilistic 3D human motion prediction via GAN". In: *CoRR* abs/1711.09561 (2017). arXiv: 1711.09561. URL: http://arxiv.org/abs/1711.09561.

[3] P. Basili et al. "Investigating Human-Human Approach and Hand-Over". In: *Human Centered Robot Systems, Cognition, Interaction, Technology*. 2009.

[4] Judith Bütepage, Hedvig Kjellström, and Danica Kragic. "Anticipating Many Futures: Online Human Motion Prediction and Generation for Human-Robot Interaction". In: May 2018, pp. 1–9. DOI: 10.1109/ICRA.2018.8460651.

[5] Zhe Cao et al. "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields". In: *CoRR* abs/1812.08008 (2018). arXiv: 1812.08008. URL: http://arxiv.org/abs/1812.08008.

[6] Enric Corona et al. "Context-Aware Human Motion Prediction". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020.

[7] T. Fong, I. Nourbakhsh, and K. Dautenhahn. "A survey of socially interactive robots". In: *Robotics and Autonomous Systems* 42.3/4 (Mar. 2003), pp. 143–166.

[8] Katerina Fragkiadaki, Sergey Levine, and Jitendra Malik. "Recurrent Network Models for Kinematic Tracking". In: *CoRR* abs/1508.00271 (2015). arXiv: 1508.00271. URL: http://arxiv.org/abs/1508.00271.

[9] Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. "Human Motion Prediction via Spatio-Temporal Inpainting". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2019.

[10] G. Hoffman and C. Breazeal. "Cost-Based Anticipatory Action Selection for Human–Robot Fluency". In: *IEEE Transactions on Robotics* 23.5 (2007), pp. 952–961. DOI: 10.1109/TRO.2007.907483.

[11] Ashesh Jain et al. "Structural-RNN: Deep Learning on Spatio-Temporal Graphs". In: *CoRR* abs/1511.05298 (2015). arXiv: 1511.05298. URL: http://arxiv.org/abs/1511.05298.

[12] Muriel Lang et al. *Object Handover Prediction using Gaussian Processes clustered with Trajectory Classification*. 2017. arXiv: 1707.02745 [cs.RO].

[13] Javier Laplaza et al. "Attention deep learning based model for predicting the 3D Human Body Pose using the Robot Human Handover Phases". In: *2021 30th IEEE International Conference on Robot Human Interactive Communication (RO-MAN)*. 2021, pp. 161–166. DOI: 10.1109/RO-MAN50785.2021.9515402.

[14] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. *History Repeats Itself: Human Motion Prediction via Motion Attention*. 2020. arXiv: 2007.11755 [cs.CV].

[15] Julieta Martinez, Michael J. Black, and Javier Romero. "On human motion prediction using recurrent neural networks". In: *CVPR*. 2017.

[16] S. Parastegari et al. "Modeling human reaching phase in human-human object handover with application in robot-human handover". In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2017, pp. 3597–3602. DOI: 10.1109/IROS.2017.8206205.

[17] Mathis Petrovich, Michael J. Black, and Gül Varol. *Action-Conditioned 3D Human Motion Synthesis with Transformer VAE*. 2021. arXiv: 2104.05670 [cs.CV].

[18] Ashish Vaswani et al. "Attention Is All You Need". In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

[19] V. Villani et al. "Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications". In: *Mechatronics* 55 (2018), pp. 248–266.