Attention deep learning based model for predicting the 3D Human Body Pose using the Robot Human Handover Phases

Javier Laplaza, Albert Pumarola, Francesc Moreno-Noguer and Alberto Sanfeliu[1]

***Abstract*— This work proposes a human motion prediction model for handover operations. We use in this work, the different phases of the handover operation to improve the human motion predictions. Our attention deep learning based model takes into account the position of the robot's End Effector and the phase in the handover operation to predict future human poses. Our model outputs a distribution of possible positions rather than one deterministic position, a key feature in order to allow robots to collaborate with humans.**

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 with an Intel RealSense D435i camera attached inside the robot's head. The results shown a great improvement of the human's right hand prediction and 3D body compared with other methods.

I. INTRODUCTION

Robot-Human interaction [8] and Robot-Human Collaboration [20] face important problems in order to be fully integrated in our society. Robots, although very common in structured and controlled environments, such as factories with working cells, are still far from being usual in unstructured environments such as houses, hospitals or offices, which are the majority of environments where humans perform their activities.

One of those important problems is to anticipate the motion of the humans around the robot. Humans can anticipate the movements of their human partners in most of daily life tasks, such as delivering tools, playing games or opening doors for example. We argue that providing robots with these skills will result in more fluent and natural human-robot interaction.

Thus, in this work we focus on the handover task between a robot and a human partner. Furthermore, we investigate the handover operation along its different phases. We divide handovers in four phases: approaching, pre-contact, contact and release, always considering that the robot is the *giver* and the human the *receiver*.

Approaching is the the first phase, when the human moves towards the robot to take the object, none of his arms is raised in a grasping position. *Pre-contact* is the next phase, where the human starts to raise his arm towards the robot. *Contact* phase begins when both the human and the robot are touching the object, followed by the *Release* phase, when the robot releases the object.

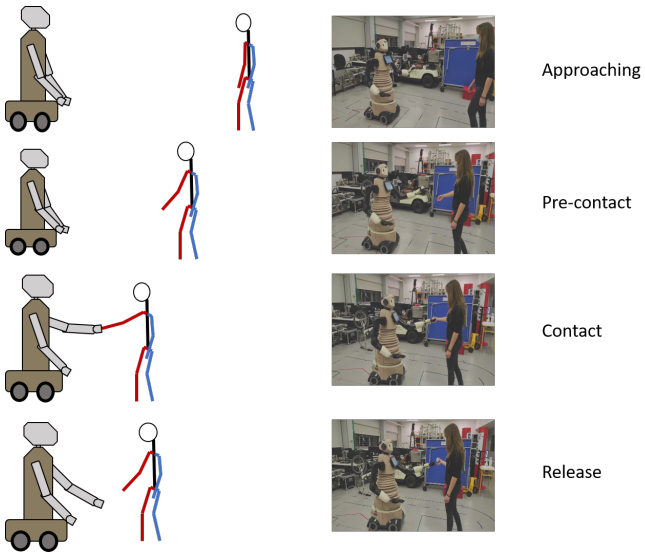

Fig. 1. We separate the handover operation in 4 different phases: approaching, pre-contact, contact and release. We provide a scheme of them

In this work we will focus on the two first phases: *Approaching* and *Pre-contact*. Since we don't pretend to tackle the grasping problem itself, we discard the *Contact* phase in our study. The *Release* is also ignored since we want to focus on the human motion, which is not relevant during this phase.

Furthermore, we believe that the most important phase during a handover operation is the pre-contact phase, since it is the phase where being able to predict the future motion of the human is most valuable. During the approaching phase, having a rough estimation of the movement of the human can be enough in order to prepare both human and robot for the handover.

For this purpose we propose an attention deep learning model that predicts the human body joint future positions while the human is approaching the robot to pick up an object that the robot has in its gripper. We compute the uncertainty of each one of the human joints and also of the human's End Effector (HEE), information that will be very valuable for the robot to adapt its plan for the handover task. The new method takes into account the robot's End Effector (REE) pose and predicts the human joint distribution instead

Work supported under the Spanish State Research Agency through the Maria de Maeztu Seal of Excellence to IRI (MDM-2016-0656), the ROCOTRANSP project (PID2019-106702RB-C21 / AEI / 10.13039/501100011033)) and the EU project CANOPIES (H2020- ICT-2020-2-101016906)

[1]Javier Laplaza, Albert Pumarola, Francesc Moreno and Alberto Sanfeliu work in the Institut de Robòtica i Informàtica Industrial in the Universitat Politecnica de Catalunya, Barcelona, Spain albert.author@papercept.net

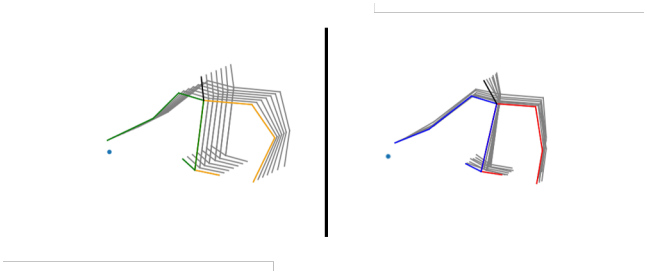

Fig. 2. Left: Model predicted trajectory. Right: Ground truth trajectory. In both cases, the blue dot shows the REE location. Both skeletons start from the same position.

of just one position of the human, as typically done in human motion prediction.

To train this model, we have created a dataset for acquiring the handover data. This dataset and the experimental validation has been done with the IVO robot shown in Fig. 1 and with a team of human volunteers that have done the picking of an object placed in the left hand of the robot.

We have made experiments with this new attention deep learning based method using the REE and the handover operation phase, and we obtained very good results that outperform the precision in the HEE and 3D human body estimation compared with other methods.

In section II we explain the related work. In section III, we present the model architecture, the goal conditioning and the probability distribution model. The dataset is explained in section IV. In section V, we describe the experimental results and finally, in section VI we present the conclusions.

II. Related work

A. Handover task

Modeling uncertainity during human-robot operations has been attempted before.

In [11], 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 [13], 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, specially nowadays since society is not yet used to be around robots. Other studies about the handover task which focus on human-human handovers are [19] and [3].

In [16], Nemlekar et al. developed an efficient method for predicting the Object Transfer Point between a robot and a human. This method is implemented into a humanoid nursing robot and experimentally validated in human-robot handover tasks. Although this method only outputs the grasping point, this method incorporates a really interesting feature to detect the human intention: gaze, a very interesting way to read the human intentionality.

In [18], Pandey et al, tackle the opposite problem, where the robot is the *taker* and the human is the *giver*. The robot must then predict where the human will handover the object to proactively try to grab the object.

B. Human Motion Prediction

Human motion prediction is a very common problem in the fields of computer vision and graphics design. Generally, motion prediction considers the human 3D skeleton as raw data. The research area can be divided in two different groups: short-term and long-term prediction. Whereas the first one focuses more in accurate prediction with little disparity with the real sequence, the second one focuses more in generating feasible human-like trajectories.

With the flourish of new deep learning techniques, new algorithms appeared to tackle this problem.

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 interesting, 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 [9], [12], [1]. The approach from [6] is specially interesting, since the model predictions are conditioned on the objects around the humans, such as tables or doors.

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 interesting 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.

The GANs have been widely used lately to propose new models. 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. To ensure that predicted human shapes look real, they add two losses to the architecture: 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 [10], GANs are used to reconstruct skeletons in sequences with occlusion problems.

To our knowledge, the closes work to our model is proposed by Diller et al. [7]. In that work, 3D poses of a skeleton are predicted considering an object placed in a table. Depending on the object type (a banana, a phone, ...) the predictions will look different. Similarly to us, they

approach the problem in a probability distribution fashion, generating a 3D grid of possible positions for each joint. The probability is generated by imposing a 3D heat-map on every joint and frame on the ground truth data and then learning that distribution, while we don't impose any prior knowledge to the distributions being generated. Furthermore, we make predictions considering that our goal, the REE, is not a static pose, since the robot's arm will move towards the human during the handover operation.

III. Model architecture

Our proposed model is inspired in the one proposed by Mao et al. [14] where we have introduced some modifications to achieve our goals. We modified the network in order to use the information of the REE inspired by the work of [17] and we also compute the pose probability distribution.

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 sequence of N human 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 poses $X_{N+1:N+T}$ representing the T future poses.

The sequence is divided into $N-M-T+1$ sub-sequences $\{X_{i:i+M+T-1}\}_{i=1}^{N-M_T+1}$, each one consisting of $M+T$.

In order to model the pose probability distribution, instead of working with the vector $X_{1:N} = [x_1, x_2, x_3, ..., x_N]$ of human joint poses, we need to define a vector $\delta X_{1:N-1} = [x_1, x_2, x_3, ..., x_{N-1}]$:

$$\delta X_{1:N-1} = X_{2:N} - X_{1:N-1} \tag{1}$$

This vector contains the difference frame by frame of the human motion, which is then discretized in B bins, considering a maximum displacement of one meter per joint.

We then tackle the problem as a classification problem by predicting which bin corresponds to the future displacement off the human per each joint. Each bin has a corresponding associated probability, which probability distribution is defined along each dimension of each joint.

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

Our approach is to create a vector $\delta = X - X_E$ to work with a gradient of distances of the human body with respect to the REE, assuming that the predictions that are generated closer to the REE will have a higher chance of raising the arm.

The motion attention allows to use the history information of the sequence in the prediction of the future sequence. The estimate is combined with the latest observed motion to be input into a Graph Convolutional Network based feed-forward network, allowing to learn the spatial and temporal dependencies in the data.

We think that the approach to identify sub-sequences can help us in the handover task, since the task involves two different phases: the approaching phase and the pre-contact phase.

The model is described as a mapping from a query and a set of key-value pairs to an output. This output is a weighted sum of values, being the weight (the attention) assigned to each value a function of the corresponding key and query.

To this end, the query, keys and REE are mapped to vectors of the same dimension *d* with three functions $f_q : \mathbb{R}^{KxM} \rightarrow \mathbb{R}^d$, $f_k : \mathbb{R}^{KxM} \rightarrow \mathbb{R}^d$ and $f_e : \mathbb{R}^{3xM} \rightarrow \mathbb{R}^d$, modeled with neural networks:

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

$$q_e = f_e(X_{E,N-M+1:N}), k_{e,i}(X_{E,i:i+M-1})$$

Where $q, k_i \in \mathbb{R}^d$ and $q^e, k_i^e \in \mathbb{R}^d$ with $i \in \{1, 2, ..., N - M - T + 1\}$. Then, the attention score is computed as follows:

$$a_i = \frac{(q + q_e)(k_i^T + k_{e,i})}{\sum_{i=1}^{N-M-T+1}(q + q_e)(k_i^T + k_{e,i})} \tag{3}$$

The model maps the resulting values to trajectory space using a Discrete Cosine Transform on the temporal dimension.

The output of the attention model is then computed as the weighted sum of values:

$$U = \sum_{i=1}^{N-M-T+1} a_i V_i \tag{4}$$

where $U \in \mathbb{R}^{k(M+T)}$. This initial estimate is then combined with the latest sub-sequence and processed by the prediction model to generate future poses $\hat{X}_{N+1:N+T}$.

To generate this probability distribution, we define a LogSoftmax final layer for the probability module.

Since the model now solves a classification problem, we use a Cross Entropy loss function, since it fits this kind of problems:

$$\mathcal{L}(\delta\hat{X}, gt) = \frac{1}{J(M+T)} \sum_{t=1}^{M+T} \sum_{j=1}^{J} -log \frac{e^{\delta\hat{X}_{t,j,gt}}}{\sum_{b=1}^{B} e^{\delta\hat{X}_{t,j,b}}} \tag{5}$$

We also add a component to penalize predictions where the final pose of the HEE is far from the REE:

$$\mathcal{L}'(\delta\hat{X}, gt) = \mathcal{L}(\delta\hat{X}, gt) + \omega_e ||\hat{p}_{M+T,rh} - p_{M+T,ee}||^2 \tag{6}$$

Where gt is the bin corresponding to the ground truth position. Once we compute the probability distribution of each joint, we sample the final pose by choosing the bin with the higher probability, and then reconstructing the whole sequence in 3D coordinates using the first frame x_1 as the starting pose:

$$p(\delta\hat{X}_{N+1:N+T}) \sim \delta X_{N+1:N+T} \rightarrow X_{N+1:N+T} \tag{7}$$

Our probability module is added to the general structure as Fig. 3 describes. This module is composed by three one dimensional convolutional layers with 512 channels, each one followed by a ReLU activation function, except the last one, which has 101 output channels (the number of bins used) and is followed by a LogSoftmax layer as described.

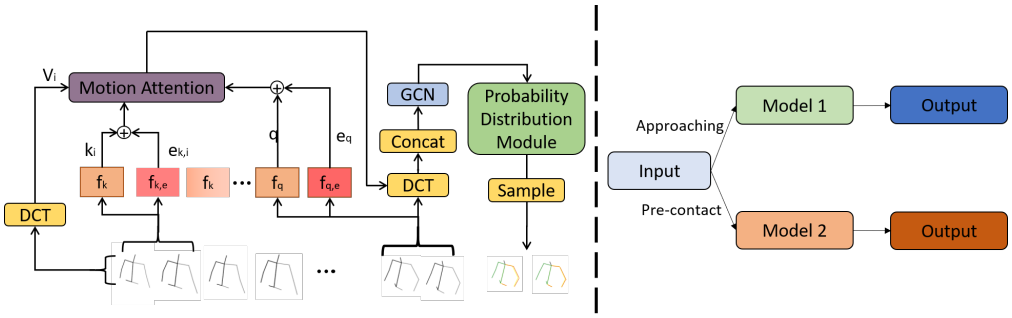

Fig. 3. Left: The model from [14] is modified with our goals: we add the $f_{k,e}$ and $f_{q,e}$ modules (red color) to condition future samples according to the distance between the skeleton and the REE. We also add the Probability Distribution Module (green color), able to generate a position distribution for each joint. We sample over this distribution to obtain the skeleton output. Right: To study each handover phase separately, we classify the input data in *Approaching* and *Pre-contact*, and train two models (Model 1 and Model 2).

IV. Dataset

A custom dataset was created consisting of human poses during a handover operation with a real anthropomorphic robot. This dataset was created to specifically study human-robot interaction during the handover task. Both the robot and the human approach towards each other and extend their arms during the sequences. The goal for the robot is to deliver a 10 cm side cube, which the human has to take from the robot's left arm End Effector. The recording ends when the human is about to separate the object from the robot gripper.

The robot used to create this dataset is the IVO robot, a humanoid robot with two arms, Fig. 1. A video from the human is recorded during the operation using an Intel RealSense D435i placed inside the robot's head. The video is recorded at a 10Hz rate. Fig. 4 shows how the sequences that were recorded from a third point perspective.

The skeleton of the human is extracted 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 had to recreate 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. Furthermore, since we are particularly interested in studying the handover operation along its different phases, each frame was labeled as *Approaching* or *Pre-contact* according to the human intention. Once the human starts raising his arm to grab the object, the sequence is considered to be *Pre-contact*. Note that the human can start raising his arm when he is still far from the robot. Also, the human can raise his arm while approaching to the robot to do other tasks such as checking his smartphone, in which case we still consider the sequence as *Approaching*.

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 scenarios possible, 15 scenarios in total, repeating once each scenario, which means 30 sequences for each volunteer, 210 sequences in total, ranging from 4 to 30 seconds. Considering that we use subsequences of 75 frames in the model and that data was visually inspected to discard corrupted data, we end up with 2.439 samples, 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.

V. Experimental Results

A. 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 perform an ablation study considering each single feature of the model separately. Since we are specially interested in the study of the different phases of the handover, we compare the model trained with all the data and the model trained only with data from each phase.

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

B. Experiments

We compute the L_2 distance in Cartesian coordinates between our predicted sequences and the ground truth se-

Fig. 4. Example of a sequence recorded for the dataset. The robot raises its left arm to offer the object to the human, who walks towards the robot too. This figure shows the sequence from a third point perspective, the video used for the dataset is recorded with the robot's head camera.

quences for the same input sequence. Table I 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.

As we can see, the mean error slightly decreases when applying the REE conditioning. However, there is a significant improvement on the human right hand prediction, when combining the pose distribution modeling and the REE conditioning.

On the other hand, we see that the mean error of the probability distribution model is very similar to the original model. This is an interesting result because this model is able to provide accurate samples with respect to the ground truth sequences, but also adds a completely new feature by modeling the probability distribution of each joint.

Finally, we train the model using only data corresponding to the *Approaching* and the *Pre-contact* phases. The outcome of this model is a significant improvement on accuracy during *Pre-contact* phase (0.1 m compared to 0.221 m). This comparison is even better for the human right hand prediction (0.073 m compared to 0.26 m). We argue that this improvement is of special interest for robot-human collaboration task, since the prediction of the human hand is specially useful during this phase.

On the other hand, the model trained only with *Approaching* data shows a similar prediction result that with all the data (0.222 m compared with 0.221 m), which isn't considered problematic since at this phase, far from the *Pre-contact*, the robot can handle the operation only knowing the average position of the whole human body. However, the accuracy on the human right is slightly improved (0.228 m compared to 0.264 m).

VI. 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 and their associated uncertainty using the information of the REE.

We used this model to study how it performs across the different phases of the handover operation, finding that the accuracy of the model is improved during the *Pre-contact* phase. We also found that the relative distance between the robot and the human can be used to improve the accuracy of the human right hand prediction, one of the most important joints to be predicted during a handover task.

One of the most useful features of the model is the ability to generate the pose distribution probability, which is required for developing human-robot shared planners for handover and other collaborative tasks.

Model	$L_2(m)$	% Samples $\leq 0.15m$	% Samples $\leq 0.25m$	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	0.207	58.67	72.78	0.349
Prob. Distr. modelling	0.224	58.78	71.21	0.365
End Effector cond + Prob. Distr. modelling	0.221	68.16	76.97	0.264
End Effector cond + Prob. Distr. modelling (Approaching)	0.222	66.35	76.18	0.228
End Effector cond + Prob. Distr. modelling (Pre-contact)	**0.100**	**85.61**	**91.5**	**0.073**

TABLE I

RESULTS OBTAINED ACROSS THE VALIDATION DATASET.

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] Christian Diller, Thomas Funkhouser, and Angela Dai. *Forecasting Characteristic 3D Poses of Human Actions*. 2020. arXiv: 2011.15079 [cs.CV].

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[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] Heramb Nemlekar, Dharini Dutia, and Zhi Li. "Object Transfer Point Estimation for Fluent Human-Robot Handovers". In: May 2019, pp. 2627–2633. DOI: 10.1109/ICRA.2019.8794008.

[17] Aaron van den Oord et al. *Conditional Image Generation with PixelCNN Decoders*. 2016. arXiv: 1606.05328 [cs.CV].

[18] A. K. Pandey et al. "Towards multi-state visuo-spatial reasoning based proactive human-robot interaction". In: *2011 15th International Conference on Advanced Robotics (ICAR)*. 2011, pp. 143–149. DOI: 10.1109/ICAR.2011.6088642.

[19] 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.

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