Context and Intention aware 3D Human Body Motion Prediction using an Attention Deep Learning model in Handover Tasks

Javier Laplaza, Francesc Moreno-Noguer and Alberto Sanfeliu

***Abstract*— This work explores how contextual information and human intention affect the motion prediction of humans during a handover operation with a social robot. By classifying human intention in four different classes, we developed a model able to generate a different motion for each intention class. Furthermore, the model uses a multi-headed attention architecture to add contextual information to the pipeline, such as the position of the robot end effector (REE) or the position of obstacles in the interaction scene. We generate predictions up to two and half seconds in the future given an input sequence of one second containing the previous motion of the human.**

The results show an improvement of the prediction accuracy, both for the full skeleton prediction and the human hand used for the delivery. The model also allows to generate different sequences with the desired human intention.

I. Introduction

Human-Robot Interaction (HRI) is a really challenging research field. While the robot side of the equation is widely researched and several improvements have been achieved, the human counterpart is a source of uncertainty during every HRI activity.

Most HRI studies relay on the human agreeing to follow a set of rules for the specific task. The human is commonly presented with a group of tools that trigger different responses from the robot. The HRI study, then evaluates how the human can create a relation with the robot using those tools. In reality, researchers have to be very clear when defining the rules to human volunteers in order to establish meaningful experiments.

When we predict how humans will move, we can't consider only the human himself, we also need to pay attention to the environment and contextual information related to the task being performed.

Nonetheless, humans will naturally move in ways that robots won't expect. In the end, these unexpected movements are the result of the human taking decisions based on his/her perception. We argue that by observing the motion of the human, robots should be able to classify their human partner attitude in order to be prepared and answer in the best possible way. Taking a further step, robots should understand how different intentions modify the human motion.

In this work we will focus on human-robot handovers. We want to study how to introduce contextual information and human intention in a human motion prediction model. Thus, our model should be able to both predict the future intention of the human and to generate different motion predictions given a desired human intention, all of this considering two contextual information queues: the position of the REE and different obstacles in the scene.

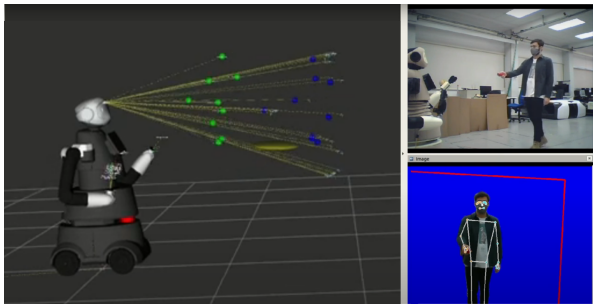

Fig. 1. Visualization of the data used by the robot during the integration. **Left:** Spatial visualization of the robot. The blue dots represent the current pose of the human. The green dots represent the predicted pose 2,5 seconds in the future. **Top-right:** Scene image seen by a external camera, for visualization purpose only. **Bottom-right:** Image recorded by the robot with the skeleton landmarks information.

II. Related Work

Contextual information has been previously used in other works that deal with motion prediction.

The approach from [3] 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 interesting work is the one presented by [14], where they use Transformer VAE, also using attention to predict the human motion, but they condition their prediction with the action that the human is performing, which arguably may 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 [11] 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 [4] or [6].

The most relevant work for our proposal is Mao et al. [10], where the temporal joint information is encoded

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

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)

using a discrete cosine transformation (DCT). This approach mitigates the problems related to auto-regressive models, and has yield to very good results in other works such as [1] by Aksan et al.

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 [5], 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 [7], Lang et al. use a Gaussian Process clustered with a stochastic classification technique for trajectory prediction using an object handover scenario. Other studies about the handover task which focus on human-human handovers are [13] and [2].

In [12], Nemlekar et al. developed an efficient method for predicting the Object Transfer Point between a robot and a human.

III. Model

We have developed a new attention deep learning model based on the Mao et al. [10], which is able to not only predict the future 3D human motion, but also the human intention.

A. Problem definition

Consider $X^p_{1:N} = [x_1, x_2, x_3, ..., x_N]$ the motion history encoding of the human motion, where $x_i \in \mathbb{R}^K$, being *K* the number of features describing each pose, in our case the 3D coordinates of each joint.

Our goal is to predict the T future poses $X^p_{N+1:N+T}$ and the predicted intention of the human for each predicted frame.

Furthermore, we want to include also contextual information related to the specific task of handover. The first contextual information we considered is the REE, since the human goal in the task is to place the object near said end effector. In consequence, we add a new queue $X^r_{1:N} = [x^r_1, x^r_2, x^r_3, ..., x^r_N]$ encoding the 3D motion history of the REE, being $x^r_i \in \mathbb{R}^3$.

The next contextual information consists on the scenario obstacles 3D position. We encode the obstacles position $X^o_{1:N} = [x^o_1, x^o_2, x^o_3, ..., x^o_N]$, where each $x^o_i \in \mathbb{R}^{3,3}$ contains the 3D coordinates of the 3 obstacles. This 3D is considered as the the obstacle centroid.

For each input sequence $X^p_{1:N}$ we also define a goal intention $i \subset [0, c-1]$ where $i \in \mathbb{N}$ and c is the number of defined intention classes (more details in Section IV). This value defines the intention that the human will express in the predicted frame $\hat{i}_{N+1:N+T}$.

B. Architecture

1) Attention channels: The first modification consists on the introduction of multiple information channels as our model input. Whereas the original model only considered the human 3D skeleton data as input, we wanted to consider multiple contextual information too. Thus, we created an attention channel for each contextual queue that we considered.

In order to compute the attention scores, we divide each input sequence $X^p_{1:N]}, X^r_{1:N}, X^o_{1:N}$ into $N - M - T + 1$ sub-sequences $X^j_{i:i+M+T-1}$, being i the time-step index of the sub-sequence and j the reference to the corresponding information channel. By creating this division, we ensure that each sub-sequence is composed by $M + T$ frames, being our goal to predict these T frames given the M previous frames. This data structure can be fit in the classical attention formulation of *keys*, *values* and *query*.

We define all the possible M length segments of the sub-sequence $X^j_{i:i+M-1}$ as the *keys*. The whole sub-sequence $X^j_{i:i+M+T-1}$ is transformed to the frequency domain using a discrete cosine transform (DCT), which output is treated as the *value* for each *key*. Finally, we take the last M frames of the sub-sequence $X^j_{N-M+1:N}$ as the *query*.

Before computing the attention scores, the keys and query are processed respectively by the mapping functions $f^j_k : \mathbb{R}^{KxM} \rightarrow \mathbb{R}^d$ and $f^j_q : \mathbb{R}^{KxM} \rightarrow \mathbb{R}^d$, which encode the input data into vectors of dimension d. Both functions are modeled using neural networks.

$$k^j_i = f^j_k(X^j_{i:i+M-1}), q^j = f^j_q(X^j_{N-M+1:N}) \tag{1}$$

2) Multi-headed Attention: In order to compute the attention scores we use multi-head attention, inspired by [15]. Basically, the same attention operation is computed in parallel inside each defined head. Each attention head receives as input a different embedding $k^{h,j}_i$ and $q^{h,j}$ for each head $h \subset [1, H]$. The attention scores for each information channel and head are then computed.

$$a^{h,j}_i = \frac{q^{h,j} k^{h,j^T}_i}{\sum_{i=1}^{N-M-T+1} q^{h,j} k^{h,j^T}_i} \tag{2}$$

3) Information fusion: The output of each attention channel is then computed:

$$U^{h,j} = \sum_{i=1}^{N-M-T+1} a^{h,j}_i V^{h,j}_i \tag{3}$$

Where each $U^{h,j} \in \mathbb{R}^{K \times (M+T)}$. This output is then concatenated with the rest of heads and fed into a linear function f_h:

$$U^j = f_h(U^{1,j} \parallel U^{2,j} \parallel ... \parallel U^{H,j}) \tag{4}$$

Finally, we perform a weighted sum of all the attention channels to obtain to obtain the attention module output:

$$U = \alpha^p U^p + \alpha^r U^r \alpha^o U^o \tag{5}$$

4) Intention conditioning: The output U is then combined with the intention conditioning module. The desired human intention is represented by i. A function $f_i : \mathbb{N} \rightarrow \mathbb{R}^{K \times (M+T)}$ is defined to map the intention information:

$$U' = U + i', \quad i' = f_i(i) \tag{6}$$

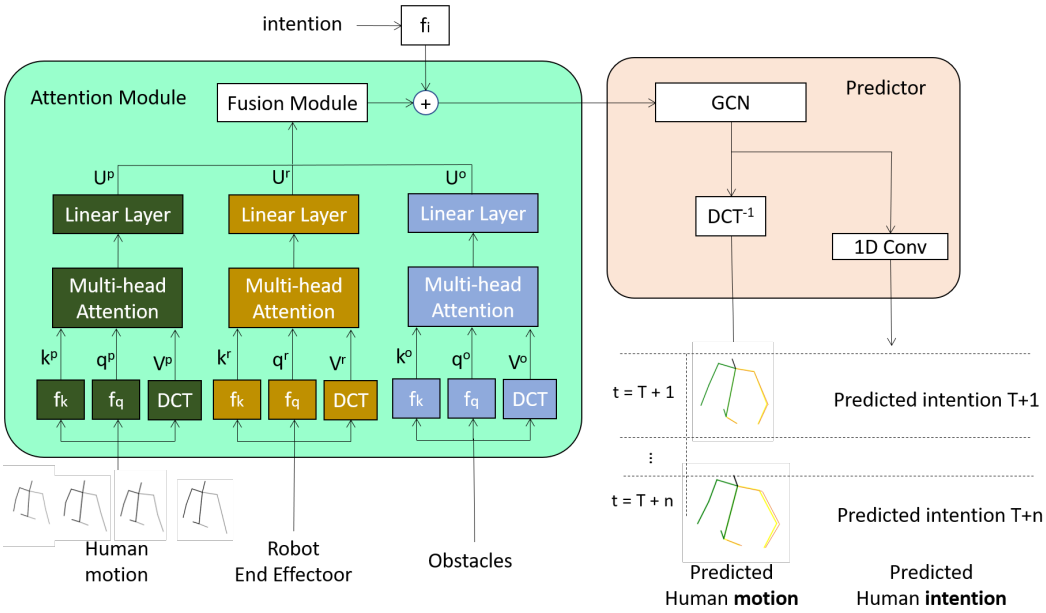

Fig. 2. Layout of the model. The left-side module corresponds to the attention architecture. The attention scores of the human motion, REE and obstacles positions are computed. An additional input representing the human intention is integrated in the module. The predictor generates both the future human motion and classifies each predicted skeleton intention.

5) Motion and intention prediction: The output U' is used by the graph convolution network (GCN) to reconstruct the predicted motion of the skeleton $\hat{X}_{N+1:N+T}$ in the same way than [10]. Additionally, we generate another output for the GCN: the predicted intention of the human for each predicted frame $\hat{i}_{N+1:N+T}$ using additional layers at the end of the GCN. These layers consist on two one-dimensional convolution layers with a ReLU activation function between them. By adding a Softmax layer at the end, we then solve a multi-class classification problem for each frame.

6) Loss function: In order to optimize our model and obtain feasible human motions, we implement several loss terms.

The main loss component is the L_2 distance between the predicted motion joints position and the ground truth position L_{xyz}.

We wanted to penalize predictions where the human hand last position is too far away from the REE since the human should try to deliver the object, thus we added L_{REE} consisting on the L_2 distance between the human right hand and the REE.

The predictions shouldn't be allowed to predict that the human will cross the obstacles of the scenario, so we added L_o to the loss to heavily penalize predictions where the human hips crossed any obstacle.

Finally, we wanted to predict the human intention in each predicted frame, so we applied a cross-entropy loss L_i in order to tackle the multi-class classification problem.

$$L = L_{xyz} + L_{REE} + L_o + L_i \tag{7}$$

IV. Dataset

In our previous work [8] we created a custom dataset in our laboratory. This time, we wanted to explore with the idea of obstacles in the scenario and the differences when the human is the one delivering the object. Thus, we created a new dataset with new conditions.

The dataset was collected using the anthropomorphic robot IVO and human volunteers performing a handover task where the human is the *giver* and the robot the *receiver* (see Fig. 3). In this case, the human takes the role of *master* and the robot takes the role of *slave*, because the robot has to follow human movements to reach the position of the object. The human and the robot approach towards each other avoiding the obstacles and extend their arms to reach their partner. At the end of this experiment, the human places the object in the robot end effector and then the robot grasp it. The delivered object is a 10 cm long cylinder handled by the human to the robot using always the right arm.

A video of each sequence is recorded using the Intel RealSense D534i camera placed inside the robot's head. The videos are recorded at 10 fps. The recording is finished when the human places the object in the REE.

The skeleton of the human is extracted from each sequence using Mediapipe [9] 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 delivers a cylindrical object to the robot in 3 different scenarios: the first scenario has no obstacles, the second scenario incorporates one obstacle between the

human and the robot and in the last scenario there are 3 obstacles. Since we wanted to have enough data representing all the different approaches that the human could take to move towards the robot, we defined different approaching paths for the humans (see Fig. 3). In the end, we defined 3 paths for the first scenario, 4 paths for the second scenario and another 4 paths for the last scenario. By creating all these situations, we wanted to study two separate aspects: how would our model responds to the human lateral movement (in our previous work we only considered straight trajectories between the human and the robot) and how would the obstacles affect our predictions.

Moreover, we ask the human volunteers to repeat three times each trajectory: the first time they are asked to perform the task in a natural way (they perform the *master - slave* behavior as expected), the second time they are asked to perform a random gesture during the task (such as waving their hands, scratch their heads, checking their smartphones, ...), although they finally deliver the object as expected, and finally they are asked to walk towards the robot and then not deliver the object (this is denominated adversarial behavior). These different behaviors were defined to allow us to study how different human intentions interfere with the motion prediction.

Once all the sequences were recorded, we performed a sanity check of the data using visual inspection. We also labeled each recorded frame with an intention class. We considered 4 different intentions: *Collaboration*, *Gesture* and *No collaboration*.

- Collaboration: the human is willing to deliver the object to the robot.
- Gesture: the human is performing a gesture (we do not differentiate between communicative and non-communicative gestures).
- Neutral: the human does not raise the right hand towards the robot, but will not make any movement to oppose the robot.
- Adversarial: the human moves the right hand away from the robot.

We also record the REE position and the robot odometry during all the sequences.

We used ten volunteers (5 women and 5 men, ages ranging from 25 to 60 years old) to perform the recordings. Each volunteer records all the possible scenarios, totaling 33 sequences for each volunteer. We end up with 330 sequences in our dataset, each sequence ranging from 4 to 15 seconds.

The human and the robot start each sequence 6 meter away from each other.

V. Training and Experimental Results

A. Training details

Since our dataset isn't very long, we decided to evaluate our model using the *leave one out* technique: we first train the model with subjects 2 to 10 and consider the human 1 as test and evaluate the accuracy of the model on the human 1 sequences, then we repeat the same but considering human 2 as test. This is repeated for all 10 humans, and we consider the average accuracy as the result.

For training, we use 50 frames (5 seconds) as input and output 25 frames(2.5 seconds). We fix the number of heads to 10, use an Adam optimizer. We perform an ablation study considering each single feature of the model separately, more the number of attention heads, the attention channels and the intention condition.

In order to compare with other methods, we train and validate other human motion prediction models in our dataset. Since we test these models in our own dataset, the results obtained might be different to the results provided in their respective papers, where they usually train their models with bigger datasets such as H3.6M and AMASS.

All the results shown in Table I are obtained using our validation dataset.

B. 3D Human motion prediction 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 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 in Table I, adding context into our predictions improve the accuracy of the model. Using the REE position information reduces the computed error of the human right hand (used to deliver the object). On top of

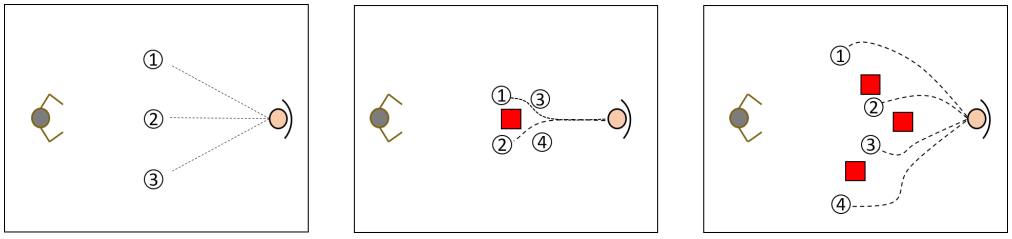

Fig. 3. Overview of the three scenarios defined in the dataset from a top-side view. For each scenario, the human is represented by the right figure, the robot is represented by the left figure and obstacles are represented by the red squares. The paths represented correspond to the human, the robot moves towards the corresponding point in each sequence.

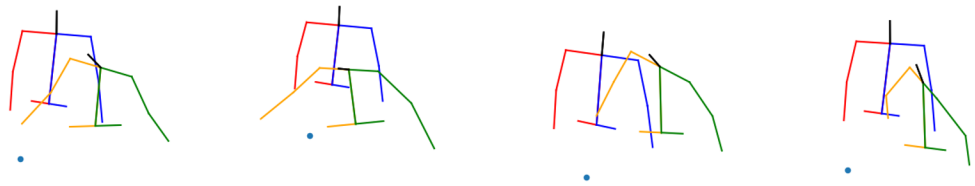

Fig. 4. Last frame of predicted sequences (green-orange) given the same input sequence using different intention goals, ground truth skeleton (red-blue) for comparison. From left to right: collaborative, gesture, neutral and adversarial. The collaborative prediction is the one where the predicted right hand position is closer to the REE (blue dot).

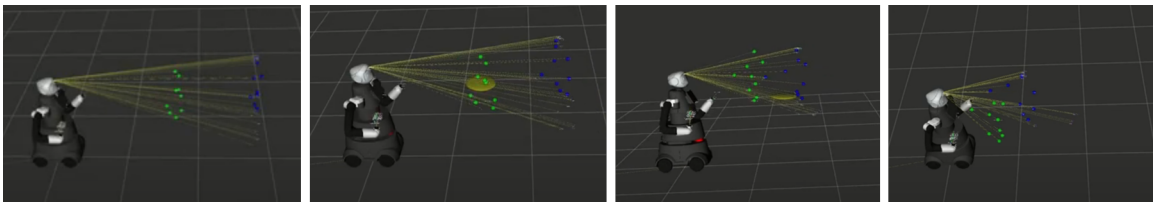

Fig. 5. Handover sequence using the prediction information in ROS environment. The blue dots represent the current position of the human, green dots represent the predicted position in the next 2.5 seconds.

that, adding the REE information also improves the accuracy of the whole upper body. A possible explanation is that, by improving the right hand accuracy, the rest of the body takes advantage using the spatial relationship between joints.

Adding the position of the obstacles seem to reduce the amount of frames with error over 35 and 40 cm. This might happen due to the skeleton being less prone to follow impossible paths and thus presenting trajectories more similar to the ground truth.

Adding the intention conditioning clearly improves the predicted intention accuracy, but the interpretation of these result can be misleading. By adding the intention conditioning in the model, we are "warning" the model with the intention of the ground truth sequence. Thus, this improvement in accuracy must be carefully considered.

Actually, by conditioning the model with the human intention we are able to generate different predicted motion based on the desired intention. Thus, given the same input sequence, we can generate one predicted motion for each human intention.

C. Handover Human-Robot validation

The predictor model was integrated in the robot as can be seen in Fig.5. The model was wrapped in a ROS node and the model feed forward was computed in a NVidia Jetson Xavier platform inside the robot. For every new message coming from the skeleton extractor, the model outputs 25 future frames representing the future 2.5 seconds.

We tested the model with two volunteers that didn't participate in the dataset collection and the results were encouraging. For the intention conditioning, we assumed that the human would collaborate with the task, but in future works we would like to use the model predicted intention for the next time step prediction.

Some model parameters had to be tuned down in order to achieve real-time performance. We used the REE to condition the predictions, but considered scenario with no obstacles.

To test the model, we repeated the task approaching the robot from different angles and observed that the predictions were consisting with the human trajectory, always facing towards the robot in the last stages.

VI. Conclusions and Future Work

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 the position information of the REE and obstacles.

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 by adding these contextual queues. Additionally, we are able to generate different predicted motions by controlling the desired intention of the prediction.

Given that the model was successfully validated in the robot, we will further study how human volunteers rate the handover interaction quality with the robot when using the prediction information.

Model	L_2 (m)	% Samples $\leq$ 0.35m	% Samples $\leq$0.40m	Right Hand L_2 (m)	Intention Accuracy
RNN [11]	0.793	3.49	11.62	0.677	-
Hist. Rep. Itself [10]	0.403	34.13	37.14	0.188	-
REE conditioning no obstacle conditioning no intention conditioning	0.378	**41.65**	45.78	0.174	56.45%
no REE conditioning obstacle conditioning no intention conditioning	0.444	41.31	44.87	0.187	62.02%
no REE conditioning no obstacle conditioning intention conditioning	0.453	30.81	36.15	0.173	86.29%
REE conditioning obstacle conditioning no intention conditioning	0.381	41.60	**47.38**	0.172	74.16
REE conditioning no obstacle conditioning intention conditioning	0.375	34.84	38.86	0.162	85.44%
no REE conditioning obstacle conditioning intention conditioning	0.387	40.22	43.73	0.17	88.69%
REE conditioning obstacle conditioning intention conditioning	**0.355**	32.15	35.73	**0.151**	**88.90**%

TABLE I

RESULTS OBTAINED ACROSS THE VALIDATION DATASET.

REFERENCES

[1] Emre Aksan et al. "Attention, please: A Spatio-temporal Transformer for 3D Human Motion Prediction". In: *CoRR* abs/2004.08692 (2020). arXiv: 2004.08692. URL: https://arxiv.org/abs/2004.08692.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] Camillo Lugaresi et al. "MediaPipe: A Framework for Building Perception Pipelines". In: *CoRR* abs/1906.08172 (2019). arXiv: 1906.08172. URL: http://arxiv.org/abs/1906.08172.

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

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

[12] 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.

[13] 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.

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

[15] 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.