Context and Intention for 3D Human Motion Prediction: Experimentation and User study in Hand-over Tasks

Javier Laplaza, Anaís Garrell, Francesc Moreno-Noguer and Alberto Sanfeliu

***Abstract*— In this work we present a novel attention deep learning model that uses context and human intention for 3D human body motion prediction in hand-over human-robot tasks. This model uses a multi-head attention architecture which incorporates as inputs the human motion, the robot end effector and the position of the obstacles. The outputs of the model are the predicted motion of the human body and the predicted human intention. We use this model to analyze a hand-over collaborative task with a robot where the robot is able to predict the future motion of the human and use this information in it's planner. We perform several experiments and ask the human volunteers to fill a standard poll to rate different features of the task when the robot uses the prediction versus when the robot doesn't use the prediction.**

I. Introduction

Most collaborative tasks between humans require a certain degree of prediction to be properly finished. Some tasks, like hand-overs, have clearly defined roles, such as master-slave, collaborative or adversary. In this kind of tasks, usually the agent acting as *slave* can predict the future motion of the other agent acting as *master* and take advantage to solve the task.

Two humans collaboratively moving a table might be an example of this, where the human leading the way is the *master* and the human following the leader is the *slave*. In this example, the *slave* can take advantage by predicting the path of the leader.

Some other interactions doesn't have strongly defined roles. Take a soccer game as example: the player with the ball will move based on his predictions of the opposing team, but also the predictions of his own teammates. At the same time, the rest of the players will predict the motion of the rest of the players. In this case, roles are not as clearly defined, but prediction of the human's body motions can severely improve the outcome of the game.

We argue that, by allowing robots to predict what humans will do in the future, robots will be able to improve the quality of human-robot interaction (HRI) tasks. We designed in this work a human motion prediction model and implemented in one of our robots (see Fig. 1). We integrated the model in the robot and use it to test how humans perceive the quality of a hand-over task when the robot is predicting the human body motion.

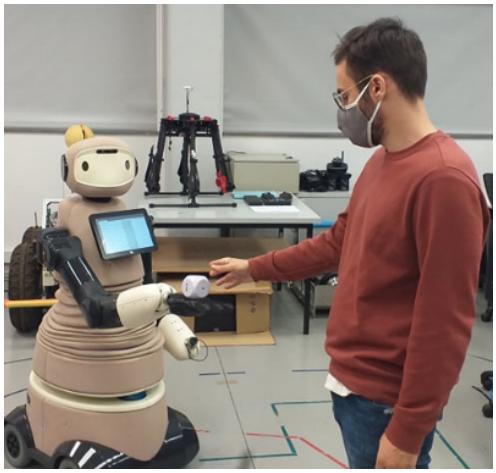

Fig. 1. We use the IVO robot to study how the human motion prediction can help during the hand-over collaborative task.

A hand-over task is defined as the action between two entities, in our case a human and a robot. The human is holding an certain object and his goal is to place this object in the robot's hand/end effector.

A hand-over task is a joint action between two agents, the giver and the receiver. The giver (in our case, the human) goal is to physically place a certain object in the receiver (in our case, the robot) hand [10].

In the reminder of the paper, in Section II we first give a short review of the related work, in Section III we explain our 3D human motion prediction model, in Section IV we describe the dataset that we created to do the validation of the model, in section V we explain the experiments, in section VI we present the user study and finally in section VII we show the conclusions.

II. Related Work

The model proposed by [3] has some similarities to ours, 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 interesting work is the one presented by [11], where they use a Transformer Variational Autoencoder (VAE) with attention architecture to predict the human motion, but the predictions are conditioned by the human action, which may be considered as context.

All authors work in the Institut de Robòtica i Informàtica Industrial de Barcelona (IRI), Catalonia, Spain jlaplaza, agarrell, fmoreno, sanfeliu@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)

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

In [9] 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 a very particular case: a non-moving skeleton can often improve results in a L2 based metric.

The most relevant work for our proposal is Mao et al. [8], where the temporal joint information is encoded 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 work proposed by [2] really lies between the purely motion prediction side and the HRI side. Butepage et al. present an unsupervised approach using VAEs to predict human motion up to 1660 ms. The work is aimed towards HRI, but the prediction model wasn't tested on a real robot.

The work of [4] presents a user study that is similar to ours, but they are focus the user study on user's security and comfort. The work also explores the prediction idea, but on the robot side. They study how to generate predictable robot trajectories (also know as "legible" trajectories) in order to improve the user experience during the collaborative task.

III. Model

We have developed a new attention deep learning model based on the Mao et al. [8], which is conditioned by the obstacles, the Robot End Effector (REE) and the human intention. The outputs of the model are the future 3D human motion and future human intention.

A. Problem definition

Let us consider $X^p_{1:N} = [x_1, x_2, x_3, ..., x_N]$ as the human motion history, 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 intention of the human for each predicted frame $\hat{i}_{N+1:N+T}$.

Furthermore, we want to include also contextual information related to the specific task of hand-over. The first contextual information we considered is the REE, since the human goal in the task is to place the object in the robot 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 second contextual information is the 3D position of the scenario obstacles. 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 N obstacles. The 3D position is 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. This value defines the intention that the human will express in the predicted frame $\hat{i}_{N+1:N+T}$. The intention classes are defined in section IV.

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 consider multiple contextual information too. Thus, we created an attention channel for each one of the contextual queues.

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 concurred with 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 [12]. Basically, the same attention operation is computed in parallel inside each defined head, which enhances the model learning capabilities. 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 as follows,

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

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:

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

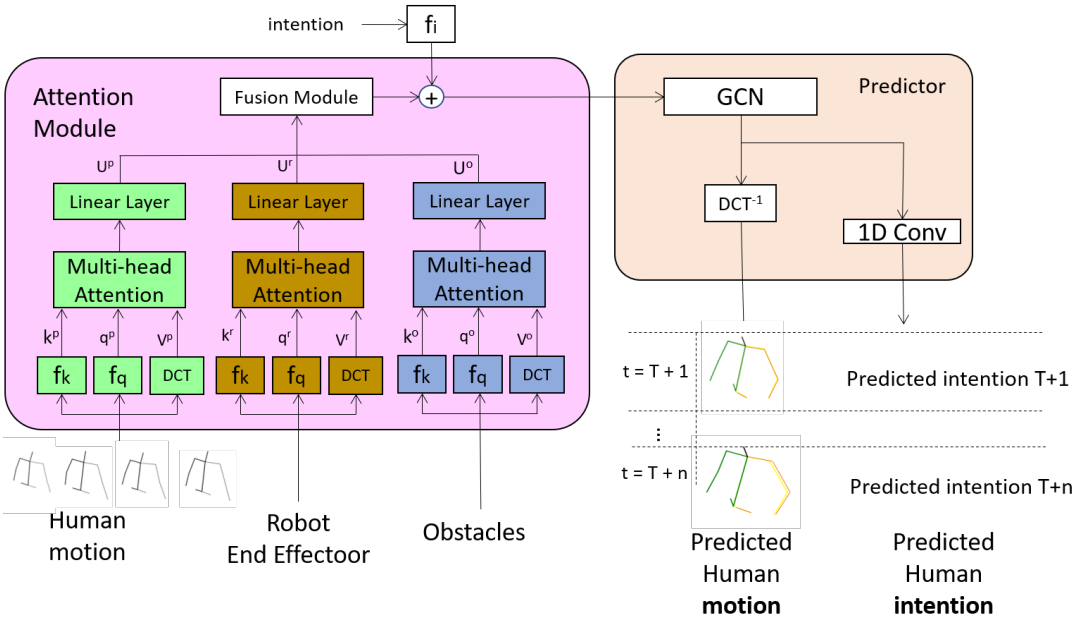

Fig. 2. Layout of the model. The left-side module corresponds to the attention architecture. The attention scores of the human motion, REE and obstacles positions are computed. An additional input representing the human intention is integrated in the module. The predictor generates both the future human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

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. This functionality will allow us to select the desired intention of the prediction that we generate each time. A function $f_i : \mathbb{N} \rightarrow \mathbb{R}^{K\times(M+T)}$ is defined to map the intention information:

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

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 [8]. 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 as the loss that heavily penalize predictions when the human hips crossed an obstacle.

Finally, we predict the human intention in each predicted frame, so we have included 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

Similarly than in our previous work [6], we created a dataset using the anthropomorphic robot IVO (Fig. 1) and human volunteers performing a hand-over 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 its arm to reach the partner. At the end of this experiment, the human places the object in the robot end effector (REE). 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 [7] to extract the 2D joint locations on the

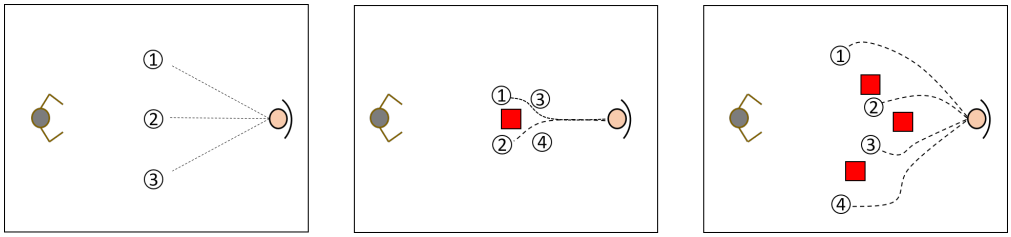

Fig. 3. Overview of the three scenarios defined in the dataset from a top-side view. For each scenario, the human is represented by the right figure, the robot is represented by the left figure and obstacles are represented by the red squares. The paths represented correspond to the human, the robot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

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 can analyze 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*, *Neutral* and *Adversarial*. This labeling process was conducted by a human watching each video frame by frame and classifying the volunteer's intention in each frame.

- 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.

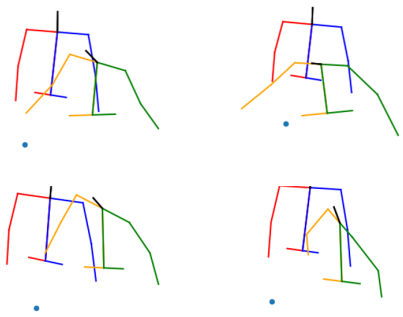

Fig. 4. Last frame of predicted sequences (green-orange) given the same input sequence using different intention goals, ground truth skeleton (red-blue) for comparison. **Collaborative** (top-left), **gesture** (top-right), **neutral**(bottom-left) and **adversarial** (bottom-right). The collaborative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

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, with an average length of 10.18 seconds and a standard deviation of 1.51 seconds.

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

V. Dataset Experiments

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.

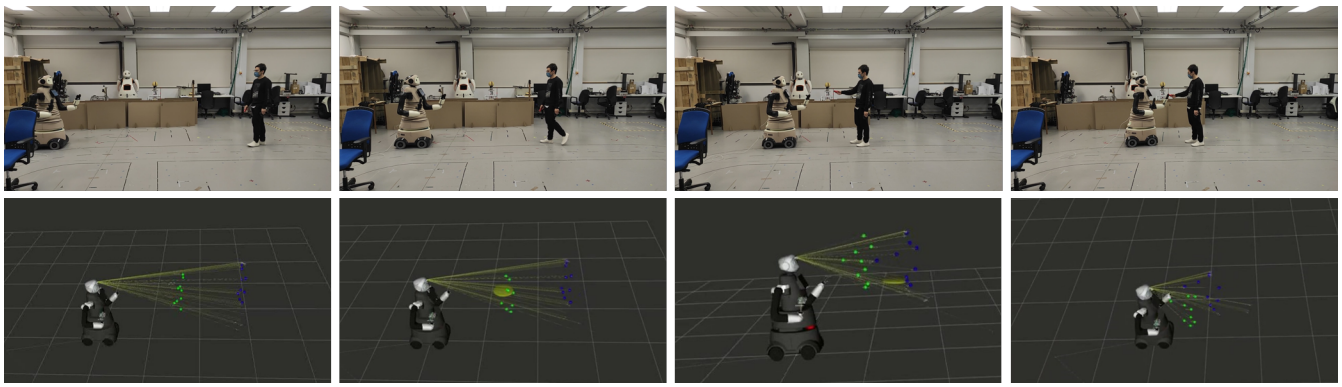

Fig. 5. Sample sequence of an experiment. The upper sequence is the RGB view of the experiment, the bottom sequence is the Rviz visualization from ROS, blue dots are the current position of the human (used in Phase B) and the green dots represent the predicted position of the human (used in Phase A).

For training, we used 50 frames (5 seconds) as input and output 25 frames (2.5 seconds). We fixed the number of heads to 10, and we used an Adam optimizer. We performed 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 trained and validated other human motion prediction models in our dataset. Since we tested 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.

B. 3D Human motion prediction experiments

We computed 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.

Finally, we checked 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 that, adding the REE information also improves the accuracy of the whole upper body.

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, as can be seen in Fig. 4.

Model	L_2 (m)	Right Hand L_2 (m)	Intention Accuracy
RNN [9]	0.793	0.677	-
Hist. Rep. Itself [8]	0.403	0.188	-
REE, no obstacle, no intention	0.378	0.174	56.45%
no REE, obstacle, no intention	0.444	0.187	62.02%
no REE, no obstacle, intention	0.453	0.173	86.29%
REE, obstacle, no intention	0.381	0.172	74.16
REE, no obstacle, intention	0.375	0.162	85.44%
no REE, obstacle, intention	0.387	0.17	88.69%
REE, obstacle, intention	**0.355**	**0.151**	**88.90%**

TABLE I

RESULTS OBTAINED ACROSS THE VALIDATION DATASET.

VI. USER STUDY

The results presented in the previous section demonstrate that the robot is able to predict and to deliver an object to a person. A user study was also conducted to determine whether the prediction module enhances the usability and the comfort of the robot from the point of view of the human.

The hypothesis we endeavored to test was as follows: "Participants will perceive a difference between the use of the prediction module and not using it."

For the experiments, we selected 15 people (8 men, 7 women) on the University Campus. Participants ranged in age from 19 to 50 years (M=29.5, SD=9.2), and represented a variety of University majors and occupations including computer science, mathematics, biology, finance and chemistry. For each individual selected, we randomly activated one of the two behaviors to deliver an object to the volunteer, it is, the activation of the prediction module or the not use of it.

For this specific experiment, we tried the model not conditioning with the REE or the obstacles position. In fact, we did not make use of obstacles in the environment, since we wanted to study purely the effect of using prediction information in the hand-over task, therefore, the presence of obstacles might interfere with the main study. Moreover, we did condition the model with the human intention, where we assumed that the human would have a collaborative attitude.

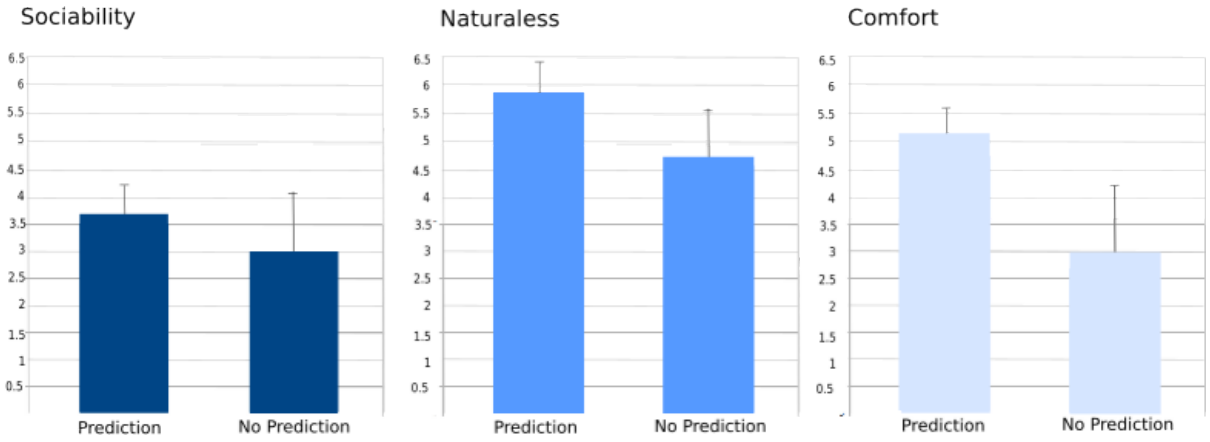

Fig. 6. Evaluation from 1 (low) to 7 (high) of the main aspects related to the robot behavior in handover task.

It should be mentioned that none of the participants had previous experience working or interacting with robots.

Participants were asked to complete a variety of surveys. Our independent variables considered whether the robot predicts humans intention or was not computing the prediction. The main dependent variables involved participants' perceptions of the **sociability**, **naturalness**, **security** and **comfort** characteristics. Each of these fields, was evaluated by every participant using a questionnaire to fill out after the experiment, based on [5].

Participants were asked to answer a questionnaire, following their encounter with the robot in each mode of behavior. To analyze their responses, we grouped the survey questions into four scales: the first measured sociability robot behavior, while the second naturalness, and third and fourth evaluated the comfort. Both scales surpassed the commonly used 0.7 level of reliability (Cronbach's alpha).

Each scale response was computed by averaging the results of the survey questions comprising the scale. ANOVAs were run on each scale to highlight differences between the three robot behaviors.

Below, we provide the results of comparing the two different behaviors. To analyze the source of the difference, four scores were examined: "sociability", "naturalness", "security" and "comfort", plotted in Fig. 6. For the sociability and security evaluation score plotted in Fig. 6, pairwise comparison with Bonferroni demonstrate there were no difference between the two kind of behavior approaches, $p = 0.3$ and $p = 0.17$, respectively . In terms of robot's naturaless and comfort, the volunteers perceived a difference between the two behaviors, $p < 0.05$ in these two cases.

Therefore, after analyzing these four components, we may conclude that if the robot is capable of predicting the human intention, then the acceptability of the robot increases.

VII. Conclusions

In this work, we propose a human motion prediction model able to use contextual information related to the hand-over task. We also condition the model with the human intention, which allow us to modify the predicted motion accordingly. The results show that the use of contextual information and intention improve the precision of the handover task.

Furthermore, we used our model in a real robot to study how do humans feel during a handover operation where the robot can predict their motion.

The experiments we conducted yielded conclusive results. We found that people felt their interaction with the robot was better overall when the robot was capable of predicting the human motion and intentions. Detailed analysis showed that this prediction improved the human's perception of the robot's naturalness, security and comfort. Hence, allowing the robot to predict its human partner seems to be appropriate for this type of scenario.

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] Judith Bütepage, Hedvig Kjellström, and Danica Kragic. "Anticipating Many Futures: Online Human Motion Prediction and Generation for Human-Robot Interaction". In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. 2018, pp. 4563–4570. DOI: 10.1109/ICRA.2018.8460651.

[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] Khoi Hoang Dinh et al. "Adaptation and Transfer of Robot Motion Policies for Close Proximity Human-Robot Interaction". In: *Frontiers in Robotics and AI* 6 (2019). ISSN: 2296-9144. DOI: 10.3389/frobt.2019.00069. URL: https://www.frontiersin.org/article/10.3389/frobt.2019.00069.

[5] Rachel Kirby. *Social robot navigation*. Carnegie Mellon University, 2010.

[6] 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.

[7] 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.

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

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

[10] Valerio Ortenzi et al. “Object Handovers: a Review for Robotics”. In: *CoRR* abs/2007.12952 (2020). arXiv: 2007.12952. URL: https://arxiv.org/abs/2007.12952.

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

[12] 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.