2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Commonsense Spatial Knowledge-aware 3-D Human Motion and Object Interaction Prediction

Sang Uk Lee[1]

***Abstract*— We propose a novel 3-D human motion and object interaction prediction model that is aware of commonsense knowledge about human–object interaction. We jointly predict human joint motion and human–object interactions. The two prediction results are combined to enforce commonsense knowledge, such as "if the human right hand is predicted to be in contact with an object after 1 second, the distance between the right hand and an object should also be predicted to be small," explicit to the model. Our model uses the raw point cloud representation of the surrounding objects in the environment as input. Using raw point cloud representation allows us to model commonsense knowledge easily and improve accuracy. In particular, it does not require a separate perception system (e.g., object classification, object pose estimation, and so on), as in previous studies, and thus is robust to perception errors. Our model applies a cross-attention mechanism to fuse the environmental point cloud and past human joint poses. The surrounding environment context and past human joint poses are two heterogeneous inputs and cross-attention can be a powerful approach to fuse them. Our model is validated on the KIT Whole-Body Human Motion (WBHM) dataset.**

I. Introduction

Human motion prediction is to predict complex 3-D motion trajectories of human joints (e.g., 18 joint human model in Figure 1a). It receives interest in various domains, including human–robot collaboration [1], computer graphics animation [2], and 3-D people tracking [3]. State-of-the-art studies make predictions solely based on past human joint history without considering environmental context [4], [5], [6], [7]. However, the environmental context is useful in predicting human motion. For example, if there is a table in a scene, human motion is constrained (or supported) by the table. Only recently, [8] proved the value of including the environmental context for 3-D human motion prediction.

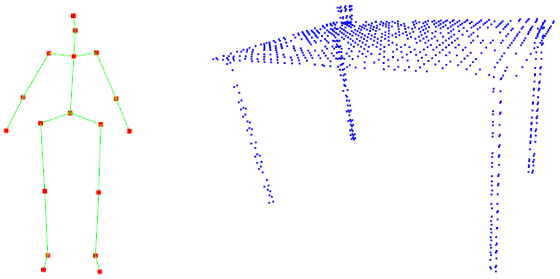
Humans understand world effectively by using commonsense knowledge. This study proposes that applying commonsense spatial knowledge about human–object interaction can improve 3-D human motion prediction significantly. For this, we perform 3-D human motion and object interaction predictions. In other words, we predict not only human joint motion, but also human–object interaction. Human–object interaction is encoded using a human-comprehensible language called qualitative spatial representation (QSR) [9]. QSR represents the spatial relations between objects using intuitive statements (e.g., "the human right hand is in contact with an object"). Using QSR, human–object interaction prediction is to predict statements such as "the human right hand is predicted to be in contact with an object after 1 second." By combining the human joint motion and QSR human–object interaction, we can easily formulate commonsense knowledge such as "if the human right hand is predicted to be in contact with an object after 1 second, the distance between the two should also be predicted small." We can make such a statement explicit to the prediction model by including a commonsense knowledge-related training loss term. It is non-trivial to make deep-learning models understand commonsense knowledge that is clear to humans [10]. We propose that this becomes easier using human-comprehensible languages such as QSR. Human-comprehensible languages were popular back in classical AI studies for encoding commonsense knowledge [11]. This study brings back an old relic to improve modern deep-learning models.

In addition to the above improvement, we apply several approaches for improving environmental context modeling for the 3-D human motion prediction. There are two crucial design factors when modeling an environmental context. The first factor is how to represent the environment [12], [13]. The second factor is how to fuse the history of human motion and the environmental context [13], [14]. The human motion history and the environmental context are heterogeneous, and how to fuse them can greatly affect the prediction accuracy. Let us briefly summarize previous context-aware prediction in [8]. Regarding the first factor, [8] modeled objects through their types (e.g., 15 types including table, cup, and so on) and bounding boxes. Regarding the second factor, [8] used graph convolution network (GCN). In the GCN in [8], each node was dedicated to a human or an object. Then, the GCN adjacency matrix captured the interaction among nodes.

We explain the two design factors in our model. Regarding the first factor, we use the raw point cloud of the surrounding objects, as shown in Figure 1b. We need not include object type because it can be inferred from the point cloud shape. A point cloud provides richer information than the bounding box. In addition, using a raw point cloud representation does not require an external perception system. In contrast, using object type and bounding box requires a separate perception system. This makes the 3-D human motion prediction system dependent on the perception system. For example, errors in perception systems (e.g., recognition misses or encountering unexpected objects) can hurt the prediction robustness. Regarding the second factor, we apply the cross-attention mechanism. The attention mechanism is the backbone of the recently developed transformer architecture [15]. Cross-attention can be powerful for fusing multi modal information [16], [17], [14], such as past human joint motion and environmental context. Our experimental results with the KIT

[1] E-mail: sang6bo@gmail.com

Whole-Body Human Motion (WBHM) dataset show that our model is more accurate and robust than previous works.



(a) A human model with 18 joints. (b) A point cloud representation of environmental context with a cup on a table.

Fig. 1: A human model with 18 joints and a point cloud representation of environmental context.

The contributions of this study are summarized as follows.

- We propose a commonsense knowledge-aware 3-D human motion and object interaction prediction model.
- We show that using point cloud representation for environment context modeling improves the accuracy and robustness to an error-prone perception system.
- We introduce a cross-attention-based 3-D human motion and object interaction prediction architecture that fuses human joint pose information and environment context.

The remainder of this study is organized as follows. Section II provides the problem statement for commonsense knowledge-aware 3-D human motion and object interaction prediction. Related works are provided in Section III. Section IV describes the proposed method. The experimental results are presented in Section V. Section VI concludes the study.

II. Problem Statement

The Goal of 3-D human motion and object interaction prediction is to predict human motions and human–object interactions from past observations. Mathematically, we construct a prediction model $\mathcal{M}$ in the following equation:

$$\mathcal{M} : \{H_{t-t_i:t}, O_{t-t_i:t}\} \rightarrow \{H_{t+1:t+t_o}, I_{t+1:t+t_o}\}. \quad (1)$$

In Eq. (1), H_t and O_t are the human and object states at time t, respectively. Then, $H_{t-t_i:t}$ and $O_{t-t_i:t}$ represent the past observation histories of human and object, respectively, from $t-t_i$ to t. Model $\mathcal{M}$ predicts $H_{t+1:t+t_o}$ and $I_{t+1:t+t_o}$ jointly. $H_{t+1:t+t_o}$ is continuous future human motion. We consider the Cartesian coordinates (i.e., x, y, and z) of 18 joints in Figure 1a. $I_{t+1:t+t_o}$ is the future human–object interaction.

We improve the prediction model by making it understand commonsense knowledge between $H_{t+1:t+t_o}$ and $I_{t+1:t+t_o}$. We formulate it mathematically as in Eq. (2) and apply it as an additional training loss term. We encode $I_{t+1:t+t_o}$ using QSR so that formulating Eq. (2) becomes easy.

$$\mathcal{L}^{CS}(H_{t+1:t+t_o}, I_{t+1:t+t_o}) \quad (2)$$

III. Related Works

A. Human Motion Prediction

Recently, deep learning has become popular in 3-D human motion prediction. Recurrent neural network (RNN) is very popular in 3-D human motion prediction [4], [5], [6]. Many works developed new feature encoding models to better extract human motion information that can be used on top of RNN architecture [18], [19], [20]. Generative adversarial models have been used to resolve the mode collapse issue (i.e., the prediction converging to a mean body pose) that many RNN-based methods suffer from [21], [22], [23], [24]. Transformer architecture has been applied to resolve rapidly accumulating errors in RNN models and allow longer prediction horizons [7], [25], [26]. These studies neglected the importance of modeling the surrounding environment context for 3-D human motion prediction. In fact, Human3.6M dataset [27], a popular human motion prediction benchmark dataset, does not contain proper annotations for the environment. Only recently, [8] improved the RNN model in [6] with environment context modeling. It experimented on the KIT WBHM dataset [28], which contains the motion capture of humans and objects with which humans interact.

B. Qualitative Spatial Representation (QSR)

QSR is a framework for representing spatial information about objects in a qualitative manner [9]. It is effective in modeling commonsense spatial knowledge [29], [30]. In particular, region connection calculus (RCC) from the QSR framework is useful for capturing mereotopological relations between objects [31]. Mereotopology is a theory of relations among the whole, parts, and boundaries. For example, "human hand is in contact with table," and "sponge is in a bowl" are RCC statements. RCC specify five possible qualitative relations: A is disconnected from B ($DC(A,B)$) (i); A is partially occluded by B ($PO(A,B)$) (ii); A is identical to B ($EQ(A,B)$) (iii); A is a proper part of B, or the inverse ($PP(A,B)$ or $PPi(A,B)$) (iv and v). Figure 2 visualizes the five relations. This study considers RCC relations between humans and the environment. For example, if a human hand is in contact with a table, it is a PO relation. If a human hand is inside a box, it is a PP relation.

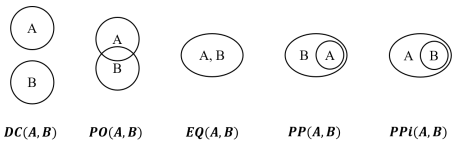


Fig. 2: RCC relations.

IV. Method

Figure 3 shows the architecture of our commonsense knowledge-aware 3-D human motion and object interaction prediction model. The model uses two point cloud inputs to predict human joint motion and RCC relations between the human and the environment. The RCC and human motion prediction results are combined through training loss function

to apply commonsense spatial knowledge. We illustrate the model in more detail in the following order: i) input, ii) architecture, and iii) loss function.

A. Inputs with Pointcloud Representation

Our model has two inputs: i) the environment context point cloud and ii) the human joint point cloud. The environment context point cloud contains the history of all the objects in the environment. Figure 3 shows an example environment context point cloud that visualizes the history of a cup moving across a table with decaying colors. The darker points represent the past. We use Fourier position encoding to encode the time flow [32], [17]. In the Fourier position encoding, we use four different frequencies and append an eight dimensional vector to each point in the point cloud. This means that the environmental context point cloud is a vector of $N_{env} \times 11$ dimensions, where N_{env} is the number of points, and 11 comes from 3 (i.e., x, y, and z) + 8. We refer to [32], [17] for details on Fourier position encoding.

We also represent the human joint history as a separate point cloud. Using point cloud for human makes sense in our work because it matches the representation for human and environment. Furthermore, [33] showed that point cloud is effective in capturing the human joint history. Figure 3 shows an example human joint point cloud. It contains a history of 18 human joints. Each joint at each time step becomes a point. We use the same Fourier position encoding to encode the time flow. For each joint at time t, we append an eight dimensional feature. Additionally, we include joint type embedding. We use a three dimensional vector for type embedding. Each dimension can have values of -1, 0, and +1. For example, we can use [-1, -1, -1] for the right elbow and [+1, 0, -1] for the left knee to distinguish them. Combining Fourier position encoding and type embedding, we append an eleven dimensional feature vector to each point in the point cloud. This implies the human joint point cloud is a vector of $N_{human} \times 14$ dimensions. The number 14 comes from 3 (i.e., x, y, and z) + 11. $N_{human} = 18 \times T_{history}$ denotes the number of points where $T_{history}$ denotes the number of time steps in the history. In our experiment, the history is 1 second long with an observation frequency of 10 Hz. Thus, we used $T_{history} = 11$, including past and current observations.

B. Model Architecture

Our model has three parts: i) PointNet [34], [35] encoder, ii) cross-attention layer, and iii) three multi-layer perceptron (MLP) heads. The PointNet encoder extracts feature from two input point clouds. We apply three PointNet layers in series to extract an $M_E \times 512$ dimensional feature for the environment context. We then apply a fully connected network (FCN) and compute an $M_E \times 256$ dimensional feature. M_E is the number of subsampled points, where $M_E = 64$. We also apply three PointNet layers and a FCN for the human joint point cloud to extract a 1×256 dimensional feature. We apply batch norm and ReLU activation layers after FCN.

After the PointNet encoder, we apply the cross-attention to fuse the two heterogeneous features. We specify the query, key, and value of the cross-attention layer. For both key and value, we concatenate $M_E \times 256$ and 1×256 dimensional features from the two point clouds. Thus, the key and value are $(1+M_E) \times 256$ dimensional vectors. We use the 1×256 dimensional feature from the human joint point cloud for the query. We apply a multi-head attention in Eq. (3) [15].

$$\begin{gathered} MultiHead(Q,K,V) = Concat(head_1, ..., head_h)W^O \\ \text{where } head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \\ \text{and } Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \end{gathered} \tag{3}$$

In Eq. (3), Q, K, and V denote the query, key, and value, respectively. $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$, and $W^O \in \mathbb{R}^{hd_v \times d_{model}}$ are the learnable matrices. This study uses eight heads; thus, $h = 8$. In addition, we use $d_{model} = 256$ and $d_k = d_v = d_{model}/h = 32$. The cross-attention computes a 1×256 feature as output.

As the final step, we apply three MLP heads to jointly predict i) the human joint poses, ii) the RCC relation between the human right hand and any object in the environment, and iii) the RCC relation between the human left hand and the environment. Each MLP head consists of two fully connected layers with a batch norm, ReLU activation, and dropout layer in the middle. The first head predicts the poses of 18 human joints over T prediction time steps. We used $T = 20$ for 2 seconds prediction horizon with a frequency of 10 Hz. The second and third heads compute RCC relations for human right and left hands, respectively, over T prediction timesteps. Computing RCC relations can be considered as a classification problem with five classes: DC, PO, EQ, PP, and PPi. Each head outputs a $5 \times T$ vector that predicts the probabilities of the five classes. We do not distinguish between object types for RCC prediction and consider all objects in the scene as one. For two cases where a human hand is in contact with i) a table and ii) a cup, we predict that the hand is in PO relation with the environment.

C. Loss Function with Commonsense Knowledge

The training loss consists of three terms as in Eq. (4).

$$\mathcal{L} = \frac{1}{I}\sum_{i}^{I}(\lambda_1 \mathcal{L}_i^{Euclidean} + \lambda_2 \mathcal{L}_i^{RCC} + \lambda_3 \mathcal{L}_i^{CS}) \tag{4}$$

In Eq. (4), λ_1, λ_2, and λ_3 are weights. We use $\lambda_1 = 1$, $\lambda_2 = 150$, and $\lambda_3 = 10$. Subscript i indicates the i^{th} training sample. Each term is described in detail below:

The first term (i.e., $\mathcal{L}_i^{Euclidean}$) denotes the mean Euclidean distance error between the predicted and target 3-D joint poses (i.e., x, y, and z). It is computed using Eq. (5).

$$\mathcal{L}_i^{Euclidean} = \frac{1}{18 \times T}\sum_{t=1}^{T}\sum_{j=1}^{18}||pose_{t,j,i}^{predict} - pose_{t,j,i}^{target}|| \tag{5}$$

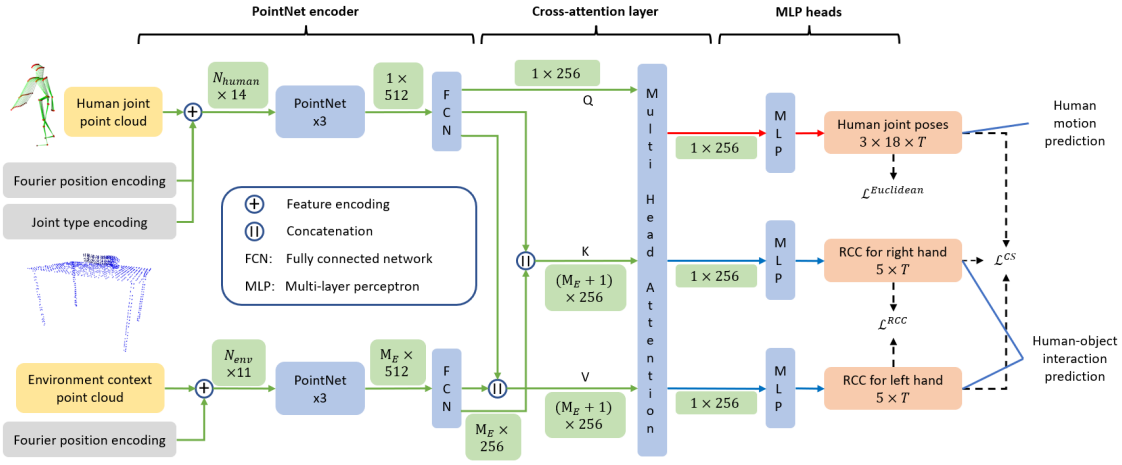


Fig. 3: 3-D human motion and object interaction prediction model architecture.

In Eq. (5), $pose^{predict}_{t,j,i}$ and $pose^{target}_{t,j,i}$ are the predicted and target joint poses, respectively. $||\cdot||$ represents the Euclidean distance. Subscripts j and t indicate the j^{th} joint and the time t, respectively.

The second term (i.e., $\mathcal{L}^{RCC}_i$) is the cross-entropy loss of RCC predictions for the human hands. RCC prediction is a classification problem with five classes: DC, PO, EQ, PP, and PPi. The second term is computed using Eq. (6).

$$\mathcal{L}^{RCC}_i = \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}^{RCC}_{t,i,right} + \mathcal{L}^{RCC}_{t,i,left} \text{ where,}$$
$$\mathcal{L}^{RCC}_{t,i,h} = \sum_{c=1}^{5} -RCC^{target}_{t,c,i,h} \times log(RCC^{predict}_{t,c,i,h}) \quad (6)$$

In Eq. (6), subscript $h \in \{right, left\}$ indicates either the right or the left hand. $RCC^{target}_{t,c,i,h}$ is 1 if class c is the correct RCC class for the sample and 0 otherwise. $RCC^{predict}_{t,c,i,h}$ is the predicted probability of class c.

The third term (i.e., $\mathcal{L}^{CS}_i$) models the commonsense knowledge about human–object interaction. It combines the prediction results of both RCC relations and human joint poses. Let us consider the following two statements: i) "if the human hand is predicted to be in contact with an object (i.e., PO, EQ, PP, or PPi), the distance between the two should also be predicted small" and ii) "if the human hand is predicted not to be in contact with all objects (i.e., DC), the distance between the hand and all objects should also be predicted large." The loss term in Eq. (7) captures these.

$$\mathcal{L}^{CS}_i = \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}^{CS}_{t,i,right} + \mathcal{L}^{CS}_{t,i,left}, \text{where}$$
$$\mathcal{L}^{CS}_{t,i,h} = \begin{cases} \max((2P^{predict}_{t,i,h}(C) - 1)(D_{t,i,h} - \epsilon_1), 0), \\ \qquad \text{if } P^{predict}_{t,i,h}(C) \geq 0.5 \\ \max((2P^{predict}_{t,i,h}(NC) - 1)(\epsilon_2 - D_{t,i,h}), 0), \\ \qquad \text{otherwise} \end{cases} \quad (7)$$

In Eq. (7), $P^{predict}_{t,i,h}(NC)$ is the predicted probability that the hand h is not in contact with any object. $P^{predict}_{t,i,h}(NC)$ is equal to $RCC^{predict}_{t,c=DC,i,h}$ from Eq. (6). $RCC^{predict}_{t,c=DC,i,h}$ is the probability of DC class that the RCC MLP head computes. $P^{predict}_{t,i,h}(C) = 1 - P^{predict}_{t,i,h}(NC)$ is the probability that the hand h is in contact with an object. $D_{t,i,h}$ is the predicted closest distance between hand h and the environmental point cloud at time t. This is computed as $D_{t,i,h} = \min_{p \in \mathcal{P}^{env}_{t,i}} ||p_{t,i,h} - p||$. $p_{t,i,h}$ is the hand joint pose at time t predicted by the first MLP head. $\mathcal{P}^{env}_{t,i}$ is the environmental point cloud at time t. We do not need to make predictions over the environmental point cloud. We use the recorded ground truth point cloud to compute the distance. This is because the distance is computed only during training, and the ground truth environmental point cloud is available during training. ϵ_1 and ϵ_2 are constant thresholds. We used $\epsilon_1 = \epsilon_2 = 180(mm)$. The first case in Eq. (7) states that if the hand is predicted to be in contact with an object and $D_{t,i,h}$ is predicted to be larger than ϵ_1 (i.e., $D_{t,i,h} > \epsilon_1$), then we penalize this. The larger $D_{t,i,h}$ is predicted to be (i.e., the further away the human hand is from any object), the more we penalize. This corresponds to the first intuitive statement described above. Similarly, the second case in Eq. (7) states that if the hand is predicted to not be in contact with any object and $D_{t,i,h}$ is predicted to be smaller than ϵ_2 (i.e., $D_{t,i,h} < \epsilon_2$), then we penalize this. The smaller $D_{t,i,h}$ is (i.e., the closer the human hand is to any object), the more we penalize. This corresponds to the second intuitive statement.

Other commonsense statements can also be considered. The distance between an object's center of mass (CoM) (e.g., a box) and a human hand pose should be small for PP and PPi relations. This can be modeled using Eq. (8).

$$\mathcal{L}^{CS,CoM}_{t,i,h} = \begin{cases} \max((2P^{predict}_{t,i,h}(In) - 1)(D^{CoM}_{t,i,h} - \epsilon_1), 0), \\ \qquad \text{if } P^{predict}_{t,i,h}(In) \geq 0.5 \\ \max((2P^{predict}_{t,i,h}(NIn) - 1)(\epsilon_2 - D^{CoM}_{t,i,h}), 0), \\ \qquad \text{otherwise} \end{cases} \quad (8)$$

TABLE I: Model Comparison. The table shows mean Euclidean distance error in *mm* and classification accuracy for human joint poses and RCC predictions respectively.

Prediction Type	Human Joint Poses (*mm*)					RCC Right					RCC Left				
Time (s)	0.5	1.0	1.5	2.0	$\sum_{t=1}^{T}$	0.5	1.0	1.5	2.0	$\frac{1}{T}\sum_{t=1}^{T}$	0.5	1.0	1.5	2.0	$\frac{1}{T}\sum_{t=1}^{T}$
ZV [6]	112.0	202.2	276.8	340.3	931.3	-	-	-	-	-	-	-	-	-	-
CA [8]	**66.2**	105.8	140.6	175.1	487.8	-	-	-	-	-	-	-	-	-	-
Raster	87.3	112.8	137.9	161.8	499.8	-	-	-	-	-	-	-	-	-	-
HMP	81.8	109.5	137.6	164.9	493.8	-	-	-	-	-	-	-	-	-	-
RCC	-	-	-	-	-	**0.937**	**0.916**	0.880	**0.863**	**0.899**	**0.914**	0.885	**0.858**	**0.842**	**0.875**
CA + RCC	68.0	106.2	139.0	169.1	482.3	0.935	0.899	**0.889**	0.858	0.895	0.903	**0.887**	0.849	0.822	0.865
HMP + RCC	76.0	**101.3**	**128.5**	**155.4**	**461.1**	0.916	0.885	0.851	0.807	0.865	0.870	0.835	0.818	0.796	0.830

TABLE II: Performance Metrics with Perception Noise.

Prediction Type	Human Joint Poses (*mm*)				
Time (s)	0.5	1.0	1.5	2.0	$\sum_{t=1}^{T}$
CA [8]	**66.2**	105.8	140.6	175.1	487.8
CA with noise	69.1	108.4	144.4	178.9	500.8
HMP + RCC	76.0	**101.3**	**128.5**	**155.4**	**461.1**

In Eq. (8), $P_{t,i,h}^{predict}(In)$ is the predicted probability that the hand h is inside an object. That is, $P_{t,i,h}^{predict}(In) = RCC_{t,c=PP \text{ or } PPi,i,h}^{predict}$. $P_{t,i,h}^{predict}(NIn) = 1 - P_{t,i,h}^{predict}(In)$ is the probability that the hand h is not inside an object. $D_{t,i,h}^{CoM}$ is the predicted distance between the hand and object CoM.

V. Experiments

We provide quantitative and qualitative evaluation results. We focus on the following two points. First, our model achieves state-of-the-art performance compared with other baselines. In particular, using commonsense knowledge about human–object interaction significantly improves human motion prediction performance. Second, other baselines that depend on a separate perception module is less robust to perception errors. Our model does not depend on a separate perception module and is thus free from such limitations.

A. Preliminaries

1) Dataset: We use the KIT WBHM dataset to validate our model following the previous context-aware human motion prediction work in [8]. We choose the KIT WBHM dataset because it not only contains human joints information but also detailed pose information of various objects humans interact with. We use 183 videos with 180K frames. It includes complex motions such as reaching an object on a table, cutting objects, mixing objects, and so on. We refer to [28], [8] for details on KIT WBHM dataset.

2) Metrics: We use the mean Euclidean distance error over 18 human joints to evaluate and compare models. The metric is computed for each prediction time step (i.e., $\mathcal{L}_t^{Euclidean} = \frac{1}{18 \times I}\sum_{i=1}^{I}\sum_{j=1}^{18}||pose_{t,j,i}^{predict} - pose_{t,j,i}^{target}||$). It is then summed over all time steps (i.e., $\mathcal{L}^{Euclidean} = \sum_{t=1}^{T}\mathcal{L}_t^{Euclidean}$). In addition, we compute the classification accuracy for models predicting RCC relation classes.

3) Baselines: We use three different baselines. First, we consider a zero-velocity (ZV) model that predicts a human to remain frozen after the last observed frame [6]. [6] showed that this simple baseline outperformed many models in certain cases. Second, we consider the context-aware human motion prediction model [8] (i.e., *CA* model). Third, we use a model shown in Figure 4 (i.e., *raster* model). We combine the two input point clouds as one. Then, we process the input with PointNet and MLP head. This model is motivated by a vehicle trajectory prediction model [36]. This work makes prediction from an input raster image that encodes the vehicle histories and surrounding environment. Human motion prediction models that only use past human motions [5], [24], [7] were not considered for comparison baselines.

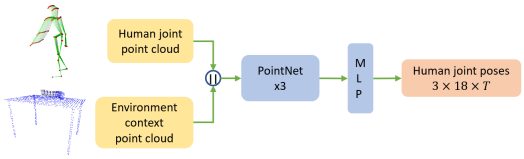


Fig. 4: *Raster* model architecture.

4) Our Models: In addition, four different models are evaluated. First, *HMP + RCC* is the model in Figure 3. Second, *HMP* includes only the human motion prediction head in Figure 3 (i.e., excluding the blue path). Third, *RCC* only predicts RCC relations (i.e., excluding the red path). Fourth, *CA + RCC* combines the baseline *CA* model and our *RCC* model. That is, we predict human joint poses with the *CA* model and RCC relations with *RCC* model. We then combine the two using the loss function in Eq. (7).

B. Quantitative Evaluations

Table I summarizes the performance metrics. All models perform better than *ZV* model. Let us first compare three deep learning models that predict human joint poses only: i) *CA*, ii) *Raster*, and iii) *HMP*. *CA* shows the best summed mean Euclidean distance error. However, the other two models also show comparable performance to *CA*.

Now, we analyze the effect of applying commonsense knowledge about human–object interaction. If we compare *HMP + RCC* and *HMP* models, we can see that it results in significant performance improvement. We can also observe a small improvement when we compare *CA + RCC* and *CA*

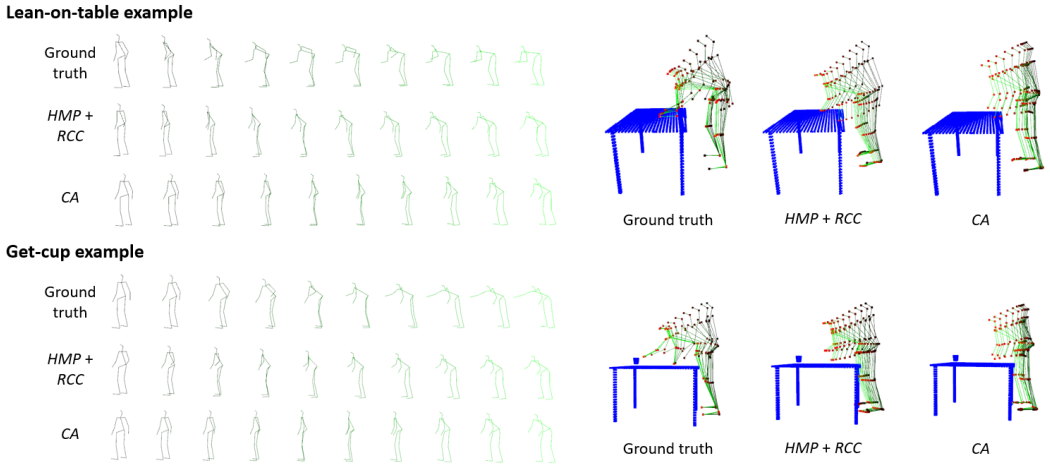


Fig. 5: 3-D human motion prediction results on lean-on-table and get-cup examples.

models. However, the improvement is more drastic when comparing *HMP + RCC* and *HMP*. We expect this is because *HMP + RCC* shares a lot of parameters for predicting the two outputs of human joint poses and RCC relations. That is, it shares the PointNet encoder and the cross-attention layer. Thus, it is more effective in learning commonsense knowledge that combines the two prediction outputs.

One interesting observation is that *CA*-based models (i.e., *CA* and *CA + RCC*) perform better in short-term predictions (i.e., $t = 0.5$). We expect this to be because of the following two points. First, RNN-based models are effective for short-term predictions, but the error quickly accumulates over long-term predictions [7], [25]. Second, simple MLP heads in Figure 3 are probably not the best architecture to decode human joint poses, especially for short-term predictions. We can use more sophisticated decoders, as in [25], [17], [6], to further improve the performance. However, the main focus of our work is on how to encode the environment context and apply commonsense knowledge rather than how to design the decoder. We achieve state-of-the-art performance even with simple MLP heads.

We also compare the RCC prediction results. If we compare *HMP + RCC*, *CA + RCC*, and *RCC* models, we can see that applying commonsense knowledge does not improve RCC prediction. We expect RCC prediction with five classes to be a relatively easier problem than human joint pose prediction. Thus, easier RCC prediction helps more difficult human motion prediction, not the other way around.

Table II summarizes the performance metrics when the models experience perception errors. *CA* baseline model requires a separate perception system to recognize the object types and bounding boxes. To mimic the perception error, we add artificial noise by removing each object from the scene with a 5% probability. The 5% error is a mild noise condition considering that state-of-the-art 3-D object detection modules show an accuracy of 90% $\sim$ 95% [34], [35]. The artificial noise does not even include errors in the object pose and bounding box estimations. *CA* undergoes about 2.7% regression in terms of the summed mean Euclidean distance error. The regression would be more severe if more noisy conditions are added. On the other hand, models using raw point cloud inputs do not experience such regression because they do not require a separate perception module.

C. Qualitative Evaluations

Figure 5 shows two example prediction results for the two models: i) *HMP + RCC* and ii) *CA*. In the lean-on-table example, the models should predict that a human will lean tightly on a table with both arms. *HMP + RCC* predicts that the human will lean more toward the table. In the get-cup example, the models should predict that a human will get a cup on a table. *HMP + RCC* predicts that the human will reach closer to the cup. In both scenarios, commonsense knowledge enforces the prediction to get closer to objects.

In the get-cup example, *HMP + RCC* predicts that the links of the human legs will cross the table. We do not consider human links when encoding commonsense knowledge. We expect that including more commonsense knowledge using human links (e.g., human links should not cross objects in the environment) could improve the results.

VI. Conclusion

This study proposes a novel 3-D human motion and object interaction prediction model. Our model uses point cloud representations of the environment and past human poses as two inputs. The two inputs are fused using a cross-attention mechanism. Our model jointly predicts human motion and human–object interactions encoded in QSR. The two predictions are combined to formulate commonsense spatial knowledge about human–object interaction.

We expect that our approach for encoding the commonsense knowledge opens the door for further research. We can include more commonsense knowledge with other high-level languages developed in classical AI research. For example, qualitative physics can be useful in encoding the kinematic and dynamic aspects of human–object interactions. Action languages can be useful for encoding the precondition and effect structure of human motion.

REFERENCES

[1] P. A. Lasota and J. A. Shah, "A Multiple-Predictor Approach to Human Motion Prediction," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 2300–2307.

[2] X. Chen, B. Jiang, W. Liu, Z. Huang, B. Fu, T. Chen, J. Yu, and G. Yu, "Executing your Commands via Motion Diffusion in Latent Space," *arXiv preprint arXiv:2212.04048*, 2022.

[3] A. Gupta, J. Martinez, J. J. Little, and R. J. Woodham, "3D Pose from Motion for Cross-view Action Recognition via Non-linear Circulant Temporal Encoding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2601–2608.

[4] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent Network Models for Human Dynamics," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4346–4354.

[5] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-RNN: Deep Learning on Spatio-Temporal Graphs," in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2016, pp. 5308–5317.

[6] J. Martinez, M. J. Black, and J. Romero, "On Human Motion Prediction Using Recurrent Neural Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2891–2900.

[7] E. Aksan, M. Kaufmann, P. Cao, and O. Hilliges, "A Spatio-temporal Transformer for 3D Human Motion Prediction," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 565–574.

[8] E. Corona, A. Pumarola, G. Alenya, and F. Moreno-Noguer, "Context-Aware Human Motion Prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6992–7001.

[9] C. Freksa, "Qualitative Spatial Reasoning," *Cognitive and Linguistic Aspects of Geographic Space*, vol. 63, pp. 361–372, 1991.

[10] N. Tandon, A. S. Varde, and G. de Melo, "Commonsense Knowledge in Machine Intelligence," *ACM SIGMOD Record*, vol. 46, no. 4, pp. 49–52, 2018.

[11] J. De Kleer and J. S. Brown, "A Qualitative Physics Based on Confluences," *Artificial intelligence*, vol. 24, no. 1-3, pp. 7–83, 1984.

[12] M. Liang, B. Yang, R. Hu, Y. Chen, R. Liao, S. Feng, and R. Urtasun, "Learning Lane Graph Representations for Motion Forecasting," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 541–556.

[13] J. Gao, C. Sun, H. Zhao, Y. Shen, D. Anguelov, C. Li, and C. Schmid, "VectorNet: Encoding HD Maps and Agent Dynamics from Vectorized Representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 525–11 533.

[14] O. Scheel, L. Bergamini, M. Wolczyk, B. Osiński, and P. Ondruska, "Urban Driver: Learning to Drive from Real-World Demonstrations Using Policy Gradients," in *Conference on Robot Learning*. PMLR, 2022, pp. 718–728.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," *Advances in neural information processing systems*, vol. 30, 2017.

[16] N. Nayakanti, R. Al-Rfou, A. Zhou, K. Goel, K. S. Refaat, and B. Sapp, "Wayformer: Motion Forecasting via Simple & Efficient Attention Networks," *arXiv preprint arXiv:2207.05844*, 2022.

[17] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira, "Perceiver: General Perception with Iterative Attention," in *International conference on machine learning*. PMLR, 2021, pp. 4651–4664.

[18] Y. Tang, L. Ma, W. Liu, and W. Zheng, "Long-Term Human Motion Prediction by Modeling Motion Context and Enhancing Motion Dynamic," *arXiv preprint arXiv:1805.02513*, 2018.

[19] E. Aksan, M. Kaufmann, and O. Hilliges, "Structured Prediction Helps 3D Human Motion Modelling," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7144–7153.

[20] A. Gopalakrishnan, A. Mali, D. Kifer, L. Giles, and A. G. Ororbia, "A Neural Temporal Model for Human Motion Prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 116–12 125.

[21] E. Barsoum, J. Kender, and Z. Liu, "HP-GAN: Probabilistic 3D Human Motion Prediction via GAN," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 1418–1427.

[22] L.-Y. Gui, Y.-X. Wang, X. Liang, and J. M. Moura, "Adversarial Geometry-Aware Human Motion Prediction," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 786–803.

[23] J. N. Kundu, M. Gor, and R. V. Babu, "BiHMP-GAN: Bidirectional 3D Human Motion Prediction GAN," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 2019, pp. 8553–8560.

[24] A. Hernandez, J. Gall, and F. Moreno-Noguer, "Human Motion Prediction via Spatio-Temporal Inpainting," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7134–7143.

[25] A. Martínez-González, M. Villamizar, and J.-M. Odobez, "Pose Transformers (POTR): Human Motion Prediction with Non-Autoregressive Transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2276–2284.

[26] W. Mao, M. Liu, and M. Salzmann, "History Repeats Itself: Human Motion Prediction via Motion Attention," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 2020, pp. 474–489.

[27] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3. 6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.

[28] C. Mandery, Ö. Terlemez, M. Do, N. Vahrenkamp, and T. Asfour, "The KIT Whole-Body Human Motion Database," in *2015 International Conference on Advanced Robotics (ICAR)*. IEEE, 2015, pp. 329–336.

[29] S. U. Lee, A. Hofmann, and B. Williams, "A Model-Based Human Activity Recognition for Human–Robot Collaboration," in *Proceedings of International Conference on Intelligent Robots and Systems (IROS)*. Macau, China: IEEE, Nov. 2019.

[30] S. U. Lee, S. Hong, A. Hofmann, and B. Williams, "QSRNet: Estimating Qualitative Spatial Representations from RGB-D Images," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 8057–8064.

[31] A. G. Cohn, B. Bennett, J. Gooday, and N. M. Gotts, "Qualitative Spatial Representation and Reasoning with the Region Connection Calculus," *GeoInformatica*, vol. 1, no. 3, pp. 275–316, Oct. 1997.

[32] K. O. Stanley, "Compositional Pattern Producing Networks: A Novel Abstraction of Development," *Genetic programming and evolvable machines*, vol. 8, pp. 131–162, 2007.

[33] Y. Zhang, M. J. Black, and S. Tang, "We are More than Our Joints: Predicting How 3D Bodies Move," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3372–3382.

[34] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[35] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space," in *Advances in neural information processing systems*, 2017, pp. 5099–5108.

[36] H. Cui, V. Radosavljevic, F.-C. Chou, T.-H. Lin, T. Nguyen, T.-K. Huang, J. Schneider, and N. Djuric, "Multimodal Trajectory Predictions for Autonomous Driving Using Deep Convolutional Networks," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 2090–2096.