WalkingDynamicsH36M: a Benchmarking Dataset for Long-term Motion and Trajectory Forecasting

Cecilia Curreli [§†‡] Andreu Girbau[†] Shin'ichi Satoh [†‡]
[§] Technical University of Munich[*] [†] National Institute of Informatics [‡] University of Tokyo

Abstract

Forecasting human pose and trajectory dynamics is a crucial step for various applications involving human-robot interaction. Compared to short-term prediction, long-term prediction can involve scenarios with stronger causal connections between past and future motions, but is a more challenging task due to the multimodal nature of the future. In this work, we present a long-term 3D prediction benchmarking dataset extracted from H36M, constraining the motion spectrum to the cyclic action of walking. We introduce a graph-based model for long-term prediction as a baseline for the task, and show limitations of currently employed distance-based metrics when evaluating the realism of predicted sequences. Our model achieves competitive results on the SoMoF dataset, predicting realistic and consistent motions on our benchmark. We believe the proposed benchmark dataset and model can serve as a foundation for future research in this field. Dataset and trained models are available on GitHub.

1. Introduction

The problem of forecasting 3D human motion is highly relevant for many realworld applications involved in human-robot interaction. This challenging problem has been approached according to different tasks, such as considering motion in terms of the sole trajectory (trajectory prediction) or in terms of the centered pose regardless of global translation (pose prediction). Recent works consider the task of predicting jointly the pose and the trajectory [1,3,5,7]. While benchmarking data already exists for short-term 3D human pose and trajectory prediction (SoMoF 3DPW [1]), to the best of our knowledge, there is no benchmarking system for long-term prediction. Depending on the scenarios, short-term predictions up to 1s may hardly capture strong dynamic motions and in particular *causal connections* between past and future motion. Considering a longer prediction time horizon allows for a deeper understanding of the *human intention* that generates future motion. The wider time window also enables us to investigate more deeply the correlation between trajectory and pose dynamics.

Long-term 3D pose and trajectory prediction faces two obstacles of stochastic nature. First, the causal connection between past and future motion is not always straightforward because of the presence of random components, especially in arm and hand motions – for instance gesticulating while speaking. Second, the same past could generate different future possibilities: while walking, a person's trajectory may bend to the left, to the right, or keep straight. These possibilities are called multi-futures, or multiple future modes.

These two probabilistic aspects make the problem of long-term pose and trajectory forecasting very challenging. We decide to constrain the motion spectrum to the cyclic action of walking.Considering only the walking action ensures an overall dynamic trajectory, eliminating quasi-static scenarios where the hip motion is close to zero or dominated by random components. This would result in an uneven real motion prior for training and in a less straightforward performance evaluation. Furthermore, a cyclic motion pattern reduces the chances of random submotion components and constrains the possible future modes to a finite number consistent over the prediction time window (compared to non-cyclic actions or arbitrary action combinations).

Our contribution is three-fold: (*i*) a benchmarking dataset WalkingDynamicsH36M with the aforementioned characteristics and suitable for multimodal prediction, (*ii*) a graph-based recurrent model GraDyn with a novel sequence parametrization as baseline for the task, and (*iii*) complementary metrics to the conventional average euclidean distance.

2. Method

We present GraDyn, a recurrent Encoder-Decoder network that relies solely on the operation of graph convolution. GraDyn is composed of TG-GRU and TG-Linear modules introduced by Motron [4]. While Motron is a multimodal method for pure pose prediction that relies on quaternion formulation, GraDyn is a deterministic model that performs

[*]This work was conducted during an internship at the National Institute of Informatics in collaboration with the University of Tokyo.

both pose and trajectory prediction with a novel sequence formulation (2.1). As in [4], the attention weights are learned according to the semantic class of each node. The encoder is a TG-GRU that maps each input sequence into a temporal representation h whereas the TG-GRU decoder decodes h into a representation g_t for each future timestep t, from which a TG-Linear layer outputs the pose and trajectory prediction $\tilde{y}_t$.

2.1. Modeling Motion Sequences

For an input sequence $x \in \mathbb{R}^{T_{\text{in}} \text{x} J \text{x} 3}$, GraDyn predicts a future sequence $\tilde{y} \in \mathbb{R}^{T_{\text{out}} \text{x} J \text{x} 3}$, where T is number of timesteps and J the number of joints. Some pose and trajectory prediction works model the motion sequence in terms of velocity, i.e. the residual between consecutive timesteps [7], or decouple the trajectory and the pose by handling them with independent networks [3]. We propose a novel parametrization of the body joints in terms of centered pose and absolute hip trajectory expressed as norm and unit vector.

Since our network is fully composed of graph convolutions, the input joints need to be mapped into graph nodes. In a pose prediction fashion, we represent each pose joint but the hip as a node and center it around the hip itself. To obtain pose nodes $n \in [-1, 1]$, we normalize the centered pose in an arbitrary cube centered around the hip, whose size depends on the chosen dataset. This approach makes the relationship between body joints straightforward and allows us to partially handle the joint prediction problem as pose prediction, while the normalization stabilizes training.

To represent the global translation or hip trajectory, we consider two different options: the velocity and the vector formulation. In the first case, we take the velocity or the residual between consecutive timesteps, mapping each hip position to a single node. While the relative velocity representation is rather limited in magnitude and therefore more stable during training, it focuses mainly on the relationship between consecutive timesteps. Considering velocities instead of coordinates is an approach commonly employed for recurrent models, and it is also applied in [7] to all joints and in [3] to the trajectory only.

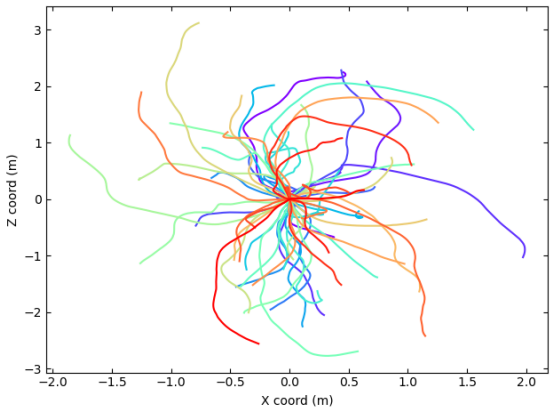

Figure 1. Hip trajectories for each test sequence of WalkingDynamicsH36M (in a different color) centered at the last input timestep.

For our vector formulation, we generate two distinct nodes from each absolute hip coordinate by splitting it into norm and unit vector. Since the unit vector and the magnitude of the global translation are absolute, this formulation keeps the correlation between non-consecutive timesteps explicit and makes the overall trajectory easier to learn in its entirety facilitating the prediction of smoother curves. These factors are particularly beneficial for learning long-term trajectories. Indeed, we saw in preliminary experiments that the vector formulation performs better in long-term prediction while the velocity formulation in short-term. In contrast, representing the trajectory as plain hip coordinates does not achieve good results as the trajectory direction is not encoded explicitly.

While the parametrization space is decoupled into centered pose and hip trajectory, the same does not hold for the learning space, as joints and trajectories are nodes of the same graph. The learned attention weights allow the network to learn suitable connections between the two spaces.

2.2. Training and Loss Functions

We train our model with two parallel losses, L_{hip} and L_{pose}. L_{hip} computes the mean squared error between prediction and ground truth sequences for the node representing the magnitude of the hip position. L_{pose} computes the mean squared error for the nodes corresponding to the centered pose and to the unit vector of the hip position. As mentioned in 2.1, all nodes involved in L_{pose} are constrained on a sphere of unit radius. Both losses are averaged over all joints and all prediction timesteps.

While averaging the mean squared error is a common loss choice in this line of works [3, 7], the formulation of these losses results from our novel joint parametrization and is motivated by the desire to balance the magnitude of the error corresponding to each node. In particular, we want to avoid imbalances originating from the hip trajectory, which generally has a higher variance compared to the other joints.

We employ curriculum learning [1] and randomly change the prediction time horizon during training as in [4] to avoid converging towards the mean of the sequences. Since L_{hip} increases with the prediction horizon, we divide it by the length of the training sequence.

Method	SoMoF 3DPW Prediction in Time					
	100 ms	240 ms	500 ms	640 ms	900 ms	**Overall**
Zero Velocity	29.4	53.6	94.5	112.7	143.1	86.7
TRiPOD [1]	31.0	50.8	84.7	104.1	150.4	84.2
SlidePose	20.2	38.1	74.4	94.9	138.8	73.3
DViTA [3]	19.5	36.9	68.3	85.5	118.2	65.7
FutureMotion [7]	9.5	22.9	50.9	66.2	97.4	49.4
SoMoFormer [5]	**9.1**	**21.3**	**47.5**	**61.6**	**91.9**	**46.3**
GraDyn (Ours)	11.8	26.7	59.0	76.8	113.6	57.8

Table 1. Results in VIM (cm) on the SoMoF 3DPW test set.

Method	MPJPE						MPJPEPose						L^2_{traj}					
	0.5 s	1 s	2 s	3 s	4 s	mean	0.5 s	1 s	2 s	3 s	4 s	mean	0.5 s	1 s	2 s	3 s	4 s	mean
DViTA [3]	41.0	434.8	999.1	1656.4	2163.6	1057.3	38.0	211.8	252.0	293.1	352.3	244.5	24.2	364.2	954.9	1598.0	2116.6	1002.3
SoMoFormer [5]	**20.6**	**280.3**	**620.8**	**1051.1**	**1417.1**	**671.6**	**22.2**	**211.1**	312.1	426.6	493.0	307.4	10.8	**221.8**	**576.3**	**1028.5**	**1374.4**	**631.9**
GraDyn (Ours)	51.6	384.5	876.0	1516.1	2078.1	964.3	52.5	217.0	**239.0**	**256.9**	**261.4**	**222.0**	**8.3**	290.1	812.8	1439.8	2013.6	891.4

Table 2. Experimental results on WalkingDynamicsH36M in mm.

Method	Prediction in Time						Joint Type		
	0.5 s	1 s	2 s	3 s	4 s	mean	std	max	min
SoMoFormer [5]	**4.3**	38.2	72.9	105.5	130.3	71.5	40.1	202.2	31.0
GraDyn (Ours)	21.0	**30.4**	**33.0**	**34.9**	**36.7**	**32.30**	**14.3**	**56.6**	**7.1**

Table 3. Limb length error on WalkingDynamicsH36M at different prediction timesteps and per joint in mm.

3. Benchmarking

3.1. WalkingDynamicsH36M

Our dataset WalkingDynamicsH36M is extracted from H36M [2] by considering only sequences that contain walking patterns sampled at 25 fps. Subjects S5 and S6 are used for validation and testing respectively, leaving the rest for training. As training data for our experiments, we densely sample as many tracks as possible from each sequence labeled as *Walking* or *WalkTogether*. For validation and testing, we select a total of 38 and 32 evaluation tracks extracted from sequences labeled with actions *Walking*, *WalkTogether*, and *Smoking* (only the walking subsequences, sampled at 16 fps). To avoid repetitive patterns, for each considered H36M sequence, we choose the frame of the first track among the first 100 randomly. Since subjects tend to walk similar paths, we randomly mirror a subset of the obtained tracks around the x-axis to evenly distribute track trajectories in space as shown in Figure 1. This homogeneous costellation makes WalkingDynamicsH36M appealing for investigating multimodality in the predictions, particularly considering multiple trajectory modes.

3.2. Metrics

We evaluate according to the conventionally employed Mean Per Joint Projection Error (MPJPE), the euclidean distance averaged over joint and prediction timesteps, and introduce two complementary metrics, MPJPEPose and L^2_{traj}. MPJPEPose measures the performance as a pure pose prediction task, computing the averaged euclidean distance after centering the predicted joints around the predicted hip position. Instead, L^2_{traj} considers only the global translation of the hip, as in trajectory prediction.

Measuring the average over long sequences does not consider the variance of the error, as the same average value can be obtained from very precise sequences with sporadic imprecision of considerable magnitude and from imprecise sequences with low variance. This is especially true as the prediction time horizon becomes longer, i.e. for long-term prediction. By decoupling pose and trajectory prediction, we seek to weaken this phenomenon and analyze the prediction from the viewpoint of other future prediction tasks — pose and trajectory prediction. These metrics allow us also to distinguish cases with similar MPJPE but opposite performance in terms of MPJPEPose and L^2_{traj}.

Metric	Baseline	Prediction in Time					
		0.5 s	1 s	2 s	3 s	4 s	mean
MPJPE	Zero Velocity	52.2	576.2	1086.3	1530.4	1860.8	1049.9
	SlidePose	36.4	355.2	810.6	1477.8	2262.6	947.7
	GT Trajectory*	35.8	198.4	240.1	275.0	330.4	230.4
L^2_{traj}	Zero Velocity	45.4	548.8	1062.1	1482.0	1801.7	1015.0
	SlidePose	7.7	276.9	740.6	1400.4	2180.4	874.0
MPJPEPose	all (Zero Vel)	37.3	206.6	250.1	286.4	344.2	240.0

Table 4. Baselines and reference measurements (*) on WalkingDynamicsH36M in mm.

Training		Prediction in Time						
T_{in}	T_{out}	0.5 s	1 s	2 s	3 s	4 s	mean@3 s	mean@4 s
1 s	4 s	24.9	319.9	692.5	1110.3	1461.6	527.3	721.0
2 s	3 s	23.9	291.7	646.6	**1027.6**	-	491.1	-
2 s	4 s	**20.6**	**280.3**	**620.8**	1051.1	**1417.1**	**477.8**	**671.6**

Table 5. Experimental results of SoMoFormer [5] on WalkingDynamicsH36M trained at different input T_{p} and output T_{f} time windows centered at the last input frames. In MPJPE (mm).

4. Experiments

4.1. Evaluation on SoMoF

We design our model for long-term prediction but evaluate it also on short-term prediction. For the experiments on SoMoF, we consider each track independently and employ the velocity formulation for the hip joint as it achieves better results. Table 1 shows that GraDyn achieves competitive results, reaching third place after the recent methods SoMoFormer [5] and FutureMotion [7]. We pretrain on 3DPW, but believe that pretraining on both 3DPW and AMASS as in higher-ranked methods [5, 7] would improve performance.

For short-term prediction, we define an algorithmic baseline, SlidePose, which slides the last input pose by the translation velocity observed between the last two input frames. As seen in Table 1, this baseline achieves competitive results on SoMoF, beating the Zero Velocity baseline and outperforming TRiPOD [1]. In our opinion, the first reason behind such good performance lies in the nature of short-term prediction, as various motions taking place in one second can be well approximated just by continuing the motion of non-static limbs along the same velocity direction. The second reason can be found in the quality of the evaluation data:

SoMoF sequences are extracted from 3DPW [6], which is collected in the wild and labeled algorithmically. This makes the data prone to noise and generates foot skating in various sequences.

4.2. Evaluation on WalkingDynamicsH36M

We evaluate recent state-of-the-art methods whose code is publicly available [3, 5] on our benchmark together with our baseline GraDyn, considering the vector formulation for the hip joint as it achieves better results.

Table 2 shows that SoMoFormer achieves the best results according to the MPJPE and L^2_{traj} metrics, while our model performs better according to MPJPEPose. In other words, SoMoFormer predicts very accurate trajectories and GraDyn predicts centered poses closer to the ground truth. SoMoFormer obtains good results for earlier prediction timesteps, since it was designed for short-term prediction, but its output does not always lie on the spectrum of real motion sequences: SoMoFormer predictions show bone deformation while our model seems to generate sequences with more realistic motion patterns (Table 3). SoMoFormer's higher bone length error might be due to it having been designed for short-term prediction where motions take place in a relatively small area, leaving less room for errors compared to longer or more dynamic sequences. Also, the original SoMoFormer implementation performs discrete cosine transformation on the input track without constraining the pose. Instead, our sequence parametrization confines limbs into a finite volume at each prediction timestep and our training optimizes these poses in a straightforward manner.

Table 4 shows further insights on WalkingDynamicsH36M. The Zero Velocity baseline achieves slightly better results than Dvita [3] while GT Trajectory*, the last seen input pose combined with the ground truth trajectory, achieves very good performance in terms of MPJPE, even if the resulting sequence is not realistic. These facts, the competitive results reached by our SlidePose baseline on SoMoF (Table 1), and the discrepancy between the MPJPE results of SoMoFormer and its bone length error, show the limitations of employing the averaged euclidean distance as a metric for realism in motion prediction.

Overall, the performance of the models evaluated on WalkingDynamicsH36M decreases over the prediction time horizon, especially according to the MPJPE and L^2_{traj} metrics, as the direction of the predicted trajectory often diverges from the ground truth. Given that long-term motion prediction has a multimodal nature, the experiment results leave open questions. Is a deterministic interpretation of the future possible, and just not learned by the models up to date, or can the task only be solved in a probabilistic manner? Is the goal of motion and trajectory prediction to generate predictions close to the ground truth or to provide realistic predictions? And what are the most suitable metrics to this end?

4.3. Discussion on WalkingDynamicsH36M

We collected WalkingDynamicsH36M as a set of sequences with a certain past time window T_{p} and future prediction horizon T_{f}. To investigate temporal causality for different temporal contexts, we consider different combinations of T_{p} and T_{f}. We define the start of each prediction time horizon as t_0 such that each sequence starts at $t = t_0 - T_{\text{p}}$ and ends at $t = t_0 + T_{\text{f}}$. We train the recent state-of-the-art model SoMoformer [5] on each of these combinations and report the evaluation results in Table 5. As is to be expected, for a prediction of $T_{\text{f}} = 4\,\text{s}$, seeing farther into the past (2 s against 1 s) lets the model achieve better prediction results as future motion *intention* is encoded not only in the immediate past. For an input time window of 2 s, SoMoFormer achieves better results when trained with a prediction horizon of 4 s instead of 3 s. This shows that seeing more of the future during training helps the model to better recognize the overall motion pattern. Based on these findings, we define WalkingDynamicsH36M with $T_{\text{p}} = 2\,\text{s}$ and $T_{\text{p}} = 4\,\text{s}$ (25 input and 100 output frames).

5. Conclusion

In this work, we introduce a novel benchmarking dataset WalkingDynamicsH36M, our model GraDyn, and show the limitations of current distance-based metrics for evaluating the realism of predicted sequences. Directions for future works can consider multi-modal future predictions.

References

[1] Vida Adeli, Mahsa Ehsanpour, Ian D. Reid, Juan Carlos Niebles, Silvio Savarese, Ehsan Adeli, and Hamid Rezatofighi. Tripod: Human trajectory and pose dynamics forecasting in the wild. In *ICCV*, 2021. 1, 2, 3

[2] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. In *PAMI*, 2014. 3

[3] Behnam Parsaeifard, Saeed Saadatnejad, Yuejiang Liu, Taylor Mordan, and Alexandre Alahi. Learning decoupled representations for human pose forecasting. In *ICCV Workshop*, 2021. 1, 2, 3, 4

[4] Tim Salzmann, Marco Pavone, and Markus Ryll. Motron: Multimodal probabilistic human motion forecasting. In *CVPR*, 2022. 1, 2

[5] Edward Vendrow, Satyajit Kumar, Ehsan Adeli, and Hamid Rezatofighi. SoMoFormer: Multi-Person Pose Forecasting with Transformers. In *arXiv preprint*, 2022. 1, 2, 3, 4

[6] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 4

[7] Chenxi Wang, Yunfeng Wang, Zixuan Huang, and Zhiwen Chen. Simple baseline for single human motion forecasting. In *ICCV Workshop*, 2021. 1, 2, 3