TRiPOD: Human Trajectory and Pose Dynamics Forecasting in the Wild

Supplementary Material

Vida Adeli[1], Mahsa Ehsanpour[2], Ian Reid[2], Juan Carlos Niebles[3],
Silvio Savarese[3], Ehsan Adeli[3], Hamid Rezatofighi[4]
[1]*Ferdowsi University of Mashhad* [2]*University of Adelaide*
[3]*Stanford University* [4]*Monash University*
http://somof.stanford.edu

A. Discussion

Why two separated H2H and H2O graphs? In this section, we discuss the reasons for considering the human to human (H2H) and human to objects (H2O) as two different graphs. *First*, these two sources of information are naturally different and the type of information and influences obtained from them are also disparate. Therefore, considering them as similar nodes of a single graph is not intuitively a sensible practice. *Second*, densely connecting these two different types of information as a single huge graph and training them all together makes it difficult for the model to converge, increases the model's complexity and the overall computation. Besides, the quality of the final features obtained are not necessarily effective. Therefore, a better practice is to consider the H2H and H2O as two different graphs but devising a solution to effectively fuse these two sources of information and their effects (described as iterative message passing in the paper).

B. Benchmark Data Details

Here, we provide more details about the two datasets that we used and re-purposed to create our human pose dynamics and trajectory forecasting benchmark.

3D Poses in the Wild (3DPW) [7]: The recently released 3DPW is a challenging outdoor dataset captured using IMU sensors, with a moving camera and consists of 60 long video clips divided into 3 train, test and validation splits. We divided the video clips into multiple non-overlapping 30-frame shorter sequences sampling over every two frames resulting in 342 sequences and to investigate the importance of predicting pose dynamics and trajectories in complex scenarios, we only consider the multi-person sequences containing social interactions. We use the 3 provided splits, However, switched the train and test splits since the number of sequences in test have become larger after the aforementioned preprocess. The body poses are in world coordinate and the results are reported in centimeter (cm). In 3DPW, the pose annotations are represented by 3D locations of 24 body joints. Since some of the joints, such as fingers and toes, are not important for the current problem, we limit our selection to a subset of 13 main body joints including the neck, shoulders, elbows, wrists, knees, hips, and ankles. In 3DPW, we feed 1000ms of past history into the model and the goal is to predict the next 1000ms of future data.

PoseTrack [2]: The PoseTrack is a large-scale multi-person dataset which covers a diverse variety of interactions including person-person and person-object in dynamic crowded scenarios. In PoseTrack, pose annotations are provided for 30 consecutive frames centered in the middle of the sequence. The pose forecasting in this dataset is challenging because of the wide variety of human actions in real-world scenarios and the large number of individuals in each sequences with large body motions and a high number of occlusions and disappearing individuals cases. Since this dataset contains cases with huge portion of joints being invisible during the time, we perform some preprocess steps to make it practicable for the current problem. We maintained only those persons that are not completely invisible in all the observation frames (means at least some partial past history should be available for a person to enable the model forecasts its future). Moreover, there were some faulty, inaccurate annotations in the dataset that we did our best to refine them. The overall number of sequences is 516 which are from the training split of this dataset. We use 60% of these sequences for training our model and the rest were split equally for validation and test. We use a set of 14 joints in 2D space defining the poses including the head, neck, shoulders, elbows, wrists, knees, hips, and ankles. The data being used is in image coordinate and therefore the results are reported in pixel. In PoseTrack sequences, we trained our model by observing the past 560ms frames and learning to minimize the prediction error over the next 560ms.

C. Input Data Types

As mentioned in the paper, we used both the offset and absolute positions as the model's input data. We practically investigated that using both offset and absolute provides the best results. The reason is that although the offset is zero mean and improves the training process, a small error in offset prediction can deviate significantly from the absolute value in high dimensions or in a long time horizon. On the other hand, the absolute is not zero mean value but keeps offset error bounded to the absolute position. Considering both information together can recompense the mutual errors.

D. Baselines Setups

The Posetrack containing invisible joints entails some initial setups for the baselines (center pose [6, 5] or trajectory forecasting[1, 3, 4]) to make it possible for them to be trained on this dataset. For training the baselines with both datasets, pose information is first centered by subtracting the neck position from every joint and the pose dynamics forecasting methods [6, 5] are trained on the local poses of the datasets. Simultaneously, the trajectory, considered as neck positions, is also learned by the three state-of-the-art trajectory forecasting methods [1, 3, 4]. Then, during prediction, we add the trajectory predictions to the local pose to obtain the global poses (results in paper, Table 1).

Moreover, to train the baselines on the PoseTrack, which contains invisible joints, we perform a similar procedure we take for training our model which means if a joint disappears from ground-truth during training, no gradient for that joint is calculated. Besides, as the neck position is required for centering the pose for pose dynamics forecasting baselines, we tried our best to refine the dataset manually, to have a good estimation of neck in occluded cases and for other cases that the agent leaves the scene we completely discard the pose. During back propagation we simply ignore these samples (do not calculate loss for them) and in test time, we use the centered poses obtained from refined neck as input and the output is whatever model predicts. Important to note that we use the refined data only for centering the pose for input and the evaluation is performed with the original data.

For the reported SC-MPF results in Table 1, we used the original SC-MPF code and metrics (requested from the authors). However, the PoseTrack data used in the SC-MPF paper is a very smaller subset of the dataset to ensure all joints for all persons in the selected sequences are fully visible as they did not model joint invisibility. We removed those assumptions from the input dataset, creating more realistic benchmarks, and used the whole dataset for the evaluation.

E. Experimental Settings

Regarding the objects used for H2O graph, we represent each object with four main features: 1) the extracted visual feature obtained from the detector 2) together with its location defined as the center location of the extracted bounding box, 3) the height and width of the bounding box, normalized over the sequence resolution and 4) the object class label as the final feature. The final object representations are obtained by passing these features through multiple MLP layers of sizes 5000, 1024 and 256. Similarly, The embedding dimensions of the MLP used for the context are 512 and 256. The hyper-parameters are selected through experiments on the validation set. We applied an initial learning rate of $5e^{-5}$ with a decay factor of 0.95 and an Adam optimizer and the step size of 2 frames being injected in each step of curriculum learning to train the model. The cut off value (β) is set to be 200 pixels. The GATs used are all single layer with 3 heads. Each experiment is performed three times and their average values are reported.

F. Additional Results

Here we provide the results for an ablation study on the number of steps performed in the iterative message passing. Table 1 shows the results. As expected, when the number of message passing iterations increase the performance first improves and then starts declining. This is commonly explored by prior graph-based learning literature [8], a crucial aspect of the graph-level representation learning is that node representations become refined and more global with the increase of the number of iterations. Therefore, it is essential to find the sufficient number of iterations for the best performance, as outlined herein.

We also investigated the effect of using a sparse or dense graph as the input skeleton representation, which is connecting the human joints in compliance with the nature of human body skeleton or representing them as fully connected graphs and letting the model to learn their relationships.The results for this study is illustrated in Table 2. The results indicate that the model can perform better when it learns the human joint relations by itself rather than sparse natural connections. This verifies the fact that the relationship between joints of an individual is not a simple hierarchical connection but every joint can have a segregated effect on each of the other joints directly.

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *CVPR*, pages 961–971, 2016. 2

[2] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele.

Table 1. Error rate for ablation study on **3DPW** dataset (in cm) using different number of message passing iterations.

Message Passing	milliseconds					
#iterations	100	240	500	640	900	AVG
1 iteration	32.49	52.71	90.39	110.51	163.46	89.91
2 iterations	32.43	52.6	89.06	109.14	159.54	88.55
3 iterations	**31.56**	**51.97**	**86.53**	**107.52**	**153.12**	**86.14**
4 iterations	33.58	52.98	91.21	111.75	163.63	90.63

Table 2. Error rate for ablation study on **3DPW** dataset (in cm) using a sparse or dense graph as input skeleton representation.

Input representation	milliseconds					
	100	240	500	640	900	AVG
Sparse Graph	33.81	53.01	89.49	110.44	158.65	89.08
Dense Graph	**31.56**	**51.97**	**86.53**	**107.52**	**153.12**	**86.14**

Posetrack: A benchmark for human pose estimation and tracking. In *CVPR*, pages 5167–5176, 2018. 1

[3] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *CVPR*, pages 2255–2264, 2018. 2

[4] Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *ICCV*, pages 6272–6281, 2019. 2

[5] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *ECCV*, 2020. 2

[6] Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *CVPR*, 2017. 2

[7] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, pages 601–617, 2018. 1

[8] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *ICLR*, 2019. 2