Effective Interaction-aware Trajectory Prediction using Temporal Convolutional Neural Networks

Noha Radwan　　Wolfram Burgard

***Abstract*— Socially compliant navigation is among the vital precursors to enable deployment of autonomous robots in populated environments. In order for mobile robots to plan trajectories that are safe and socially acceptable, they need to accurately and efficiently predict future trajectories of the surrounding pedestrians. Although current approaches for motion prediction utilize data-driven methods to estimate the behavior of pedestrians, they only model a local neighborhood around each pedestrian. This often results in suboptimal behavior in densely populated environments, where the interactions are more complex in nature. In this work, we propose the IA-TCNN architecture to address the problem of efficient trajectory prediction for multiple interacting pedestrians in a scene. Our network leverages the previous motion information of all pedestrians in order to aggregate trajectories of the most relevant pedestrians within the scene, thus predicting an accurate future trajectory. Extensive experimental evaluations on indoor and outdoor benchmarks demonstrate that our approach achieves state-of-the art performance, while simultaneously achieving fast inference time thereby facilitating online deployment.**

I. Introduction

Among the major goals of robotics is the development of intelligent platforms that are capable of performing a variety of tasks in everyday life for their users. Over the previous decade, robots have become more integrated into our daily lives; performing numerous tasks including domestic cleaners, navigational aids and last-mile delivery. In particular for mobile robots that share the space with humans, compliant navigation is a crucial capability. Identifying and fulfilling such behavior in itself is a skill that we learn as humans over several years and we reiterate this learning process whenever we go to a different society. Hence, hard-coding a set of behavioral rules for a mobile robot to abide by is not only tedious, but also requires constant upkeep depending on the environment. Recently, learning-based motion prediction approaches [1], [4], [5] have shown considerable robustness in modeling pedestrian interactions in various environments. However as the complexity of the scene increases, the run-time and representational capabilities of such approaches substantially decreases, since they rely on modeling each pedestrian separately by considering only their local neighborhood.

In this work, we propose a novel scalable approach to address the problem of learning trajectories in populated environments. We frame the problem of trajectory estimation as a sequence-to-sequence modeling task. We utilize a data-driven method to represent the behavior of pedestrians, thus enabling our approach to leverage the inherent interdependencies in the motion thereby learning interactions without manually specifying a set of behavioral rules [2], [7]. Instead of the widely employed recurrent units such as LSTMs, our proposed network employs causal convolutions which facilitates both accurate modeling of the sequential behavior of the pedestrians and online deployment in resource constrained systems. In order to evaluate the applicability of our approach we perform exhaustive experiments in a diverse set of indoor settings and outdoor environments across different cities.

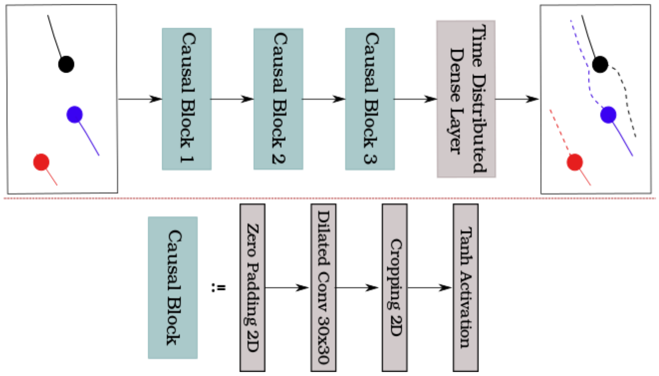


Fig. 1. Schematic representation of our proposed IA-TCNN architecture for motion prediction. Input to the network is the observed trajectories for all pedestrians over a time interval. For each pedestrian, the network predicts the trajectory for the upcoming interval by capturing the interactions among the various pedestrians in the scene.

II. Technical Approach

Given the trajectory information for each pedestrian over a certain time period, the output of our model is the corresponding trajectory for each pedestrian over the prediction interval. We define the trajectory $\mathcal{O}_i$ for pedestrian i during an observation interval $T_{\text{obs}} = \{1, \ldots, t_{\text{obs}}\}$ as:

$$\mathcal{O}_i = \left\{ \left(x_i^t, y_i^t, qw_i^t, qz_i^t\right) \in \mathbb{R}^4 \mid t \in T_{\text{obs}} \right\}, \qquad (1)$$

where each trajectory point is represented by the spatial coordinates (x_i^t, y_i^t) and the yaw angle $q_i^t = (qw_i^t, qz_i^t)$ in normalized quaternion representation. Our network produces the predicted trajectory $\mathcal{W}_i$ over the interval $T_{\text{pred}} = \{t_{\text{obs}} + 1, \ldots, t_{\text{pred}}\}$ such that:

$$\mathcal{W}_i = \left\{ \left(x_i^t, y_i^t, qw_i^t, qz_i^t\right) \in \mathbb{R}^4 \mid t \in T_{\text{pred}} \right\}. \qquad (2)$$

In order to represent this problem as a sequence-to-sequence modeling task, the predicted output at timestep

All authors are with the Institute of Computer Science, University of Freiburg, Germany.

$t \in T_{\text{pred}}$ can only depend on inputs from $t/ \in T_{\text{obs}}$. In other words, predictions cannot depend on future states of traffic participants. Moreover, we predict the future trajectories for an interval greater than or equal to the observation interval, as estimating the trajectories for an interval shorter than the observation interval is rather simple. In this work, however, our goal is to accurately predict the future trajectories of pedestrians for intervals longer than the observation intervals.

We propose the *Interaction-aware Temporal Convolutional Neural Network (IA-TCNN)*, depicted in Fig. 1, to address the above requirements. Our network consists of three causal blocks; where each block contains zero-padding followed by dilated causal convolution, cropping and `tanh` activation. In each block, we employ zero padding and cropping layers to satisfy the requirement of predicting a trajectory with length greater than or equal to the observed trajectory. We utilize causal convolutions where the output at each timestep is convolved with elements from earlier timesteps, thereby preventing information leak across different layers. Although the amount of previous information utilized by causal convolutions is linear to the network depth, increasing the depth or using extremely large filter sizes increases the inference time as well as the training complexity. We overcome this problem by employing dilated causal convolutions to increase the receptive field without increasing the depth of the network. We use a constant kernel size of 30×30 for each of the convolutional layers with filter sizes of $[128, 128, 128]$ respectively, and increase the dilation rate by 1 for each following block. We model the predicted spatial coordinates of each pedestrian using a bivariate Gaussian distribution to obtain a measure of confidence over the output of the network. The output of the last block is passed to a time distributed dense layer of size 7 to produce temporal predictions for each timestep of the prediction interval, where for each pedestrian the network predicts the mean $\mu_i^t = (\mu_x, \mu_y)_i^t$, standard deviation $\sigma_i^t = (\sigma_x, \sigma_y)_i^t$, correlation coefficient ρ_i^t, and quaternion q_i^t.

We train our model by minimizing the weighted combination of the negative log likelihood loss of the groundtruth position under the predicted Gaussian distribution parameters and the $\mathcal{L}_2$ loss of the orientation in normalized quaternion representation as follows:

$$\begin{aligned} \mathcal{L}_\gamma &= \left\| q_i^t - \hat{q}_i^t \right\|_2 \\ \mathcal{L}_p &= -\log\left(\mathbb{P}\left(x_i^t, y_i^t \mid \hat{\mu}_i^t, \hat{\sigma}_i^t, \hat{\rho}_i^t\right)\right) \\ \mathcal{L}_{MP} &= \sum_i^N \sum_t^{t_{\text{pred}}} \mathcal{L}_p \exp(-\hat{s}_p) + \mathcal{L}_\gamma \exp(-\hat{s}_\gamma) + \hat{s}_p + \hat{s}_\gamma, \end{aligned} \tag{3}$$

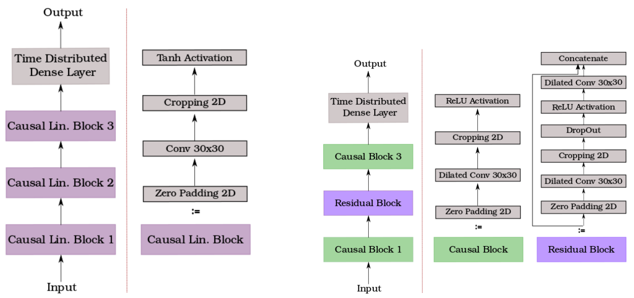
where N is the number of pedestrians, $\hat{s}_p$, $\hat{s}_\gamma$ are learnable weighting variables for balancing the translational and rotational components of the predicted pose.

Since in real world data the trajectories of the different pedestrians have varying lengths due to the limited sensor range, and in order to fully leverage all the information available during training, we train our proposed IA-TCNN with dynamic sequence lengths by using binary activation masks predicted by the network to signify the end of a trajectory. This in turn implicitly enables the network to learn when a pedestrian exits the field of view of the sensor. The predicted trajectory is then first multiplied by the activation mask before computing the prediction error. Moreover, as opposed to explicitly selecting for each pedestrian the set of pedestrians likely to affect its behavior, our proposed model utilizes information from all pedestrians during the observation interval to predict the trajectory for each of the observed pedestrians. This has the advantage of eliminating the need for creating handcrafted definitions which attempt to explicitly model how the behavior of a pedestrian is affected by surrounding pedestrians. Furthermore, it expedites the information flow throughout the various layers of the network, hence facilitating fast trajectory estimation for all pedestrians in the scene.

III. Experimental Evaluation

We evaluate the performance of our proposed IA-TCNN on both the indoor L-CAS dataset [8] and the outdoor ETH [3] and UCY [2] crowd set datasets. The L-CAS dataset is composed of over 900 pedestrian tracks divided into a training and a test split. Each pedestrian track has an average length of 13.5s, wherein each pedestrian is identified by a unique ID, time frame at which they were detected, spatial coordinates and orientation angle. There are several factors that make benchmarking on this dataset extremely challenging such as people pushing trolleys and children running, in addition to groups forming and dispersing.

The ETH crowd set dataset consists of two scenes: Univ and Hotel, containing a total of approximately 750 pedestrians exhibiting complex interactions. Each tracked pedestrian is identified by a pedestrian ID, frame number and spatial coordinates at which they were observed. Similar to the ETH dataset, the UCY dataset is a crowd set dataset comprised of three scenes: Zara01, Zara02 and Uni, with a total of approximately 780 pedestrians. For each scene, the dataset provides an annotations file consisting of a series of splines each describing the trajectory of a pedestrian using the spatial coordinates, frame number and the viewing direction of the pedestrian. This dataset in addition to the ETH dataset are widely used in conjunction as benchmarks for motion prediction and pedestrian tracking due to the wide range of non linear trajectories and pedestrian interactions exhibited. In order to facilitate comparison with existing work on these benchmarks, we use the provided train and test split for the L-CAS dataset, while we combine both the ETH and UCY datasets, similar to previous works [1], [6], and apply a leave-one-out procedure during training, by randomly selecting trajectories from all scenes except the testing scene. Furthermore, for the L-CAS dataset, we utilize and predict the spatial and angular information for each pedestrian, while for the ETH and UCY datasets, we only predict the



(a) IA-LinConv (b) IA-DResTCNN

Fig. 2. Schematic representation of two variants of our proposed architecture; IA-LinConv and IA-DResTCNN.

2D spatial coordinates of each pedestrian to facilitate the combination of the data.

In order to train our IA-TCNN model such that it is robust to the varying number of pedestrians observable in every interval, we introduce a variable to represent the maximum number of distinct trajectories observed within an interval and initially set it to the maximum observed in all datasets. During training and testing, we use an activation mask to encode the positions of valid trajectories and discard all remaining information. We train our approach for 100 epochs with a mini-batch size of 12. We employ the Adam solver for optimization, with a learning rate of 0.0005 and apply gradient clipping. All experiments are conducted on the Tensorflow library on a single Nvidia Titan X GPU.

We evaluate the accuracy of our motion prediction model by reporting the following metrics:

- *Average Displacement Error*: mean squared error over all predicted and groundtruth points in the trajectory.
- *Final Displacement Error*: the distance between the predicted and groundtruth poses at the end of the prediction interval.

In order to evaluate the efficacy of our proposed convolutional network for the sequence modeling task, we create two variants of our method; namely IA-LinConv and IA-DResTCNN, depicted in Fig. 2. IA-LinConv closely resembles IA-TCNN with the exception of employing standard convolutions in place of the dilated convolutions. While in the IA-DResTCNN, we replace the middle causal block by a residual block, and the tanh activation function by the standard ReLU activation.

Following the evaluation procedure of [5], we train on a sequence length of 20 frames on the L-CAS dataset, and use an observation and prediction of 8 and 12 frames respectively during testing. The average displacement error of our approach in comparison to current state-of-the-art methods is shown in Tab. I. Both our baseline approaches, IA-LinConv and IA-DResTCNN, are able to outperform the standard recurrent-based methods by 64.2%, and 32.0% in the translational and rotational components respectively which in turn corroborates the advantage of utilizing a convolutional architecture over recurrent methods. Moreover, by utilizing our proposed IA-TCNN, we are able to achieve an average displacement error of 0.11m and 21.7° further improving upon the achieved results by 67.6% and 8.8% in translation and orientation respectively. This improvement over the results achieved by IA-LinConv is attributed to employing dilated convolutions which increase the receptive field, thereby increasing the amount of information utilized at each layer. However, we observe that adding a residual block to our network as in IA-DResTCNN did not help in improving the prediction accuracy over IA-TCNN.

TABLE III
Comparison with the state-of-the-art on UCY-Uni.

Method	Avg. Disp. Error (m)	Final Disp. Error (m)	Run-time (s)	Size (MB)
Social-LSTM [1]	0.27	0.77	1.78	95.8
IA-TCNN (Ours)	0.29	**0.46**	**0.26**	**7.0**

In Tab. II, we demonstrate the average displacement error of our proposed approach in comparison to state-of-the-art methods on the different sequences of the ETH and UCY datasets. We train our model using a sequence length of 20 observations, and during testing we observe 8 frames (corresponding to 3.2s) and predict the upcoming 12 frames (4.8s). Note that for each of the methods, we report the numbers directly from the corresponding manuscripts, with the exception of the Social Forces model where we report the numbers from [1] as the original manuscript does not report the same metrics as the ones employed by the state-of-the-art methods. Utilizing the proposed architecture, we achieve an improvement of 29.6% in comparison to the previous state-of-the-art. Despite the sparse amount of sequences available for these datasets, and the complexity of the pedestrian interactions demonstrated, our method is able to achieve the lowest final displacement error, as depicted in Tab. IV, with an improvement of 55.7% in comparison to previous works on all the sequences.

Tab. V shows the effect of varying the observation and prediction lengths of the average displacement accuracy of our proposed IA-TCNN approach on the Uni sequence of the UCY crowd set dataset. For short observation lengths ($2-4$ frames), the error in the predicted trajectory linearly increases with the increase in the prediction length. This accounts for the increased difficulty of making accurate predictions given short trajectory information as future interactions cannot be reliably predicted. Concurrently, by increasing the observation length, the prediction accuracy gradually increases with small improvements between $6-8$ observation frames. This can be attributed to the reduction in the amount of significant information over time due to the short interaction times between pedestrians and the low likelihood of abrupt changes in the behavior of one or more pedestrians.

We further compare the run time and model size of our approach with Social-LSTM [1] in Tab. III. The results show that using our proposed IA-TCNN, we improve upon the final displacement accuracy by 40.3% while being 85.4% faster than Social-LSTM [1]. Moreover, our proposed approach only requires 7.0MB of storage space, which is

TABLE I

AVERAGE DISPLACEMENT ACCURACY OF IA-TCNN IN COMPARISON TO EXISTING METHODS ON THE L-CAS DATASET.

Dataset	Social-LSTM [1]	Pose-LSTM [5]	IA-LinConv	IA-DResTCNN	IA-TCNN (Ours)
L-CAS	1.19m, NAN	0.95m, 35.0°	0.34m, 23.8°	0.46m, 33.1°	**0.11**m, **21.7**°

TABLE II

AVERAGE DISPLACEMENT ACCURACY OF IA-TCNN ON THE ETH AND UCY DATASETS IN COMPARISON TO STATE-OF-THE-ART METHODS.

Dataset	Social Forces [7]	Basic LSTM	Social-LSTM [1]	Social-Attention [6]	IA-LinConv	IA-DResTCNN	IA-TCNN (Ours)
ETH-Univ	0.41m	0.39m	0.50m	0.39m	0.27m	0.43m	**0.15**m
ETH-Hotel	0.25m	0.32m	0.11m	0.29m	0.28m	0.36m	0.16m
Zara01	0.40m	0.18m	0.22m	0.20m	0.34m	0.45m	**0.14**m
Zara02	0.40m	0.28m	0.25m	0.30m	0.38m	0.37m	**0.19**m
UCY-Uni	0.48m	0.30m	0.27m	0.33m	0.41m	0.36m	0.29m
Average	0.39m	0.29m	0.27m	0.30m	0.34m	0.39m	**0.19**m

TABLE IV

FINAL DISPLACEMENT ACCURACY OF IA-TCNN IN COMPARISON TO EXISTING METHODS ON THE ETH AND UCY DATASETS.

Dataset	Social Forces [7]	Basic LSTM	Social-LSTM [1]	Social-Attention [6]	IA-LinConv	IA-DResTCNN	IA-TCNN (Ours)
ETH-Univ	0.59m	1.06m	1.07m	3.74m	0.27m	0.60m	**0.21**m
ETH-Hotel	0.37m	0.33m	0.23m	2.64m	0.32m	0.52m	**0.18**m
Zara01	0.60m	0.93m	0.48m	0.52m	0.54m	1.08m	**0.27**m
Zara02	0.68m	1.09m	0.50m	2.13m	0.47m	0.88m	**0.25**m
UCY-Uni	0.78m	1.25m	0.77m	3.92m	0.66m	1.03m	**0.46**m
Average	0.60m	0.93m	0.61m	2.59m	0.45m	0.82m	**0.27**m

TABLE V

EFFECT OF THE VARYING THE OBSERVATION AND PREDICTION LENGTHS IN FRAMES ON THE AVERAGE DISPLACEMENT ERROR FOR OUR PROPOSED IA-TCNN METHOD ON THE UCY-UNI DATASET.

Obs. Length \ Pred. Length	2	3	4	5	6	7	8	9	10	11	12
2	0.42m	0.48m	0.50m	0.53m	0.56m	0.61m	0.63m	0.62m	0.62m	0.61m	0.60m
3	0.31m	0.36m	0.41m	0.45m	0.49m	0.52m	0.52m	0.51m	0.51m	0.51m	0.51m
4	0.23m	0.30m	0.35m	0.39m	0.42m	0.43m	0.43m	0.43m	0.43m	0.43m	0.44m
5	0.20m	0.26m	0.31m	0.36m	0.36m	0.37m	0.37m	0.37m	0.38m	0.38m	0.38m
6	0.18m	0.23m	0.28m	0.29m	0.30m	0.32m	0.32m	0.33m	0.33m	0.33m	0.34m
7	0.15m	0.20m	0.22m	0.24m	0.26m	0.27m	0.28m	0.29m	0.29m	0.30m	0.31m
8	0.14m	0.17m	0.19m	0.21m	0.23m	0.24m	0.25m	0.26m	0.26m	0.27m	0.29m

substantially lesser than its counterparts, enabling it to be efficiently deployed in resource limited systems.

IV. CONCLUSION & FUTURE WORK

In this paper, we presented an approach for interaction-aware motion prediction using a temporal convolutional architecture which accurately predicts the trajectory of pedestrians. Our approach efficiently encodes observations from all pedestrians in the scene, thereby rendering it scalable to complex environments while simultaneously predicting accurate trajectories. Extensive experimental evaluations demonstrate that our IA-TCNN achieves state-of-the-art performance on both indoor and outdoor datasets, while achieving faster inference times in comparison to recurrent approaches. Regarding future work, we aim to additionally predict the obstacle map of the environment, as we believe that knowledge about the vicinity can improve the overall prediction accuracy by avoiding trajectories that are occupied by obstacles.

REFERENCES

[1] A. Alahi *et al.*, “Social lstm: Human trajectory prediction in crowded spaces,” in *CVPR*, 2016.

[2] A. Lerner, Y. Chrysanthou, and D. Lischinski, “Crowds by example,” in *Computer Graphics Forum*, vol. 26, no. 3. Wiley Online Library, 2007, pp. 655–664.

[3] S. Pellegrini *et al.*, “You’ll never walk alone: Modeling social behavior for multi-target tracking,” in *ICCV*, 2009.

[4] M. Pfeiffer *et al.*, “A data-driven model for interaction-aware pedestrian motion prediction in object cluttered environments,” in *ICRA*, 2018.

[5] L. Sun *et al.*, “3dof pedestrian trajectory prediction learned from long-term autonomous mobile robot deployment data,” in *ICRA*, 2018.

[6] A. Vemula, K. Muelling, and J. Oh, “Social attention: Modeling attention in human crowds,” in *ICRA*, 2018.

[7] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, “Who are you with and where are you going?” in *CVPR*, 2011.

[8] Z. Yan, T. Duckett, and N. Bellotto, “Online learning for human classification in 3d lidar-based tracking,” in *IROS*, 2017.