Intent-Aware Human Motion Prediction using Deep Generative Neural Networks

Kapil D. Katyal[1,2], I-Jeng Wang[1], Gregory D. Hager[2] and Chien-Ming Huang[2]

***Abstract*— A critical capability required for the wide adoption of mobile robots into society is the ability to navigate safely around pedestrians. One important component to enable safe navigation is to accurately predict the motion of pedestrians in the scene. The main objective of this research is to develop novel techniques that accurately predict human motion by using past motion and intent as a prior for making the prediction. In this study, we develop neural network architectures that are capable of learning environment-agnostic embeddings that serve as a prior for prediction. We combine these embeddings with contextual information including desired velocity and a probability distribution describing the intent to make predictions. We compare the average displacement error and final displacement error with state-of-the-art published results and show evidence that combining contextual information results in more accurate prediction of future motion.**

I. Introduction

For decades, we have envisioned a world where humans and robots interact seamlessly in society. An important component of this interaction is to allow the robot to navigate safely around pedestrians. Humans are extremely capable of exhibiting this skill and seamlessly navigate around other humans or groups of humans. We postulate that predicting future trajectories of surrounding pedestrians is an important characteristic that enables this seamless navigation skill. In this paper, we present an algorithm that enables the prediction of human trajectories in space.

Specifically, our main contribution is a framework that combines learning a latent representation using deep neural networks with contextual information. We have demonstrated that our approach, when combined with contextual information such as desired velocity and target goals can lead to better long term prediction accuracy. We compare our approach to algorithms that only learn a latent representation from a neural network without using explicit contextual information.

II. Related Work

The study of human motion prediction has been researched significantly in the past. In a work by Karasev et al., the authors describe an approach to long term motion prediction by modeling the pedestrian behavior using jump-Markovian processes where the intent of the pedestrian is modeled as a latent variable [1]. Another recently published work by Gupta et al. describe an approach that combines a generative adversarial network, a recurrent neural network and novel pooling mechanism to aggregate information across individuals or pedestrians [2]. A more comprehensive survey on existing taxonomies, approaches, prior art and future work has been studied by Rudenko et al. [3]. Our approach differs from other approaches by leveraging contextual information in addition to embeddings learned by the deep neural network.

[1]Johns Hopkins University Applied Physics Lab, Laurel, MD, USA. Kapil.Katyal@jhuapl.edu

[2]Dept. of Comp. Sci., Johns Hopkins University, Baltimore, MD, USA.

III. Problem Definition

In this section, we formally define the problem of pedestrian motion prediction. Similar to [2], we model the state of pedestrian i as $X_i = (x_i^t, y_i^t)$, where $t = 1, ..., t_{obs}$ and predict future trajectories as $\hat{Y}_i = (x_i^t, y_i^t)$ where $t = t_{obs} + 1, ..., t_{pred}$. In our experiments, we used values of $t_{obs} = 8$ and $t_{pred} = 8$ and 12 corresponding to a prediction of 3.2 and 4.8 seconds, respectively.

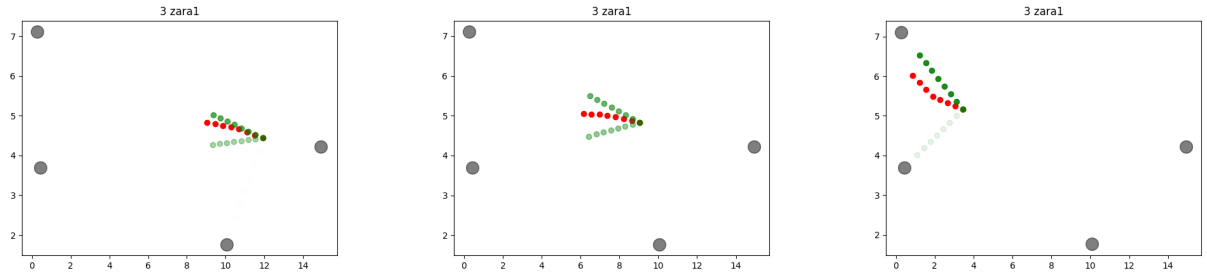
IV. Experiment Details

A. Architecture

Our architecture for prediction consists of an autoencoder based neural network that consists of encoder, decoder and hidden layers. Our encoder function uses a combination of linear fully connected layers and an LSTM layer to capture time series data. Our hidden layer includes the bottleneck of the autoencoder along with combined data that explicitely captures contextual information described below. The decoder network also consists of an LSTM decoding layer and fully connected layer. The activation functions for all layers include a Rectified Linear Unit. The input of our network is t_{obs} $\times$ batch_size $\times$ state_size, where the state_size = 2 representing the (x_i^t, y_i^t) coordinates. The output of the neural network is t_{pred} $\times$ batch_size $\times$ state_size. The network is trained using mean squared error between the predicted trajectories and the ground truth as the loss function.

B. Contextual Information

In this preliminary study, we investigate two sources of contextual information that can be considered as a prior for prediction. The first is the average velocity of the pedestrian. It seems straightforward that the average velocity of the pedestrian will serve as a strong prior for predicting future motion and it is not obvious that the LSTM network of the neural network will capture this. The second contextual information that we considered is the intended target or goal of the pedestrian. In this work, we cluster the final positions of the pedestrian using a small portion of the dataset to find discrete goals. During training and testing, we estimate a probability distribution of the goals using a Gibbs distribution.



(a) Start of the Trajectory (b) Middle of the Trajectory (c) End of the Trajectory

Fig. 1. This figure demonstrates our ability to estimate the long term goal of the human pedestrian. The red circles represent the observed positions of the pedestrian. The green circles represent the idealized trajectories to each of the target positions. The opacity of the green colored circles represents the probability of the observed trajectory mapping to goal trajectory. The darker circles represent a higher probability. (Best viewed in color)

C. Datasets and Metrics

Our experiments use two highly cited publicly available datasets for training and evaluation purposes : ETH [4] and UCY [5]. These two datasets consist of 5 datasets representing 4 scenes in crowded, outdoor environments. The metrics used for evaluation are consistent with other research in this space [1], [2] and include the average displacement error (ADE) and the final displacement error (FDE). The ADE is the average mean squared error between the predicted positions and the ground truth over the entire predicted trajectory. The FDE is the mean squared error between the final position of the predicted trajectory and the ground truth.

V. Preliminary Results

In Fig. I, we show a few examples of our ability to estimate the probability of the desired goal using a Gibbs distribution. The red circles represent the observed positions of the pedestrian during the observation period. The green circles represent the idealized trajectories from the start of the observed trajectory to each of the target positions. The opacity of the green colored circles represents the probability of the observed trajectory mapping to goal trajectory.

Tables I and II compare the ADE and FDE in meters between our approach using only the velocity information and SocialGAN [2]. The values in the table represent the results using t_{pred} = (8 / 12). In most cases, we are able to observe an increased accuracy when adding the contextual information.

TABLE I
Average Displacement Error

Dataset	SGAN [2]	Ours
ETH	0.58 / 0.71	0.58 / **0.61**
HOTEL	0.36 / 0.48	**0.30** / **0.36**
UNIV	0.33 / 0.56	0.33 / **0.51**
ZARA1	0.21 / 0.34	0.21 / **0.30**
ZARA2	0.21 / 0.31	**0.19** / **0.26**
Average	0.34 / 0.48	**0.32** / **0.41**

TABLE II
Final Displacement Error

Dataset	SGAN [2]	Ours
ETH	**1.13** / **1.29**	1.15 / 1.33
HOTEL	0.71 / 1.02	**0.56** / **0.75**
UNIV	0.70 / 1.18	**0.69** / **1.10**
ZARA1	0.42 / 0.69	**0.41** / **0.60**
ZARA2	0.42 / 0.64	**0.39** / **0.56**
Average	0.68 / 0.96	**0.64** / **0.87**

VI. Conclusion and Future Directions

To successfully integrate mobile robots into our society to provide various assistance, it is critical to ensure that they can navigate safely around humans. In this paper, we explore how deep generative neural networks can be used for effective human motion prediction. We present preliminary results describing the benefits to combine contextual information with latent representations learned from a deep neural network for motion prediction. There are many directions we plan to take this research. We plan to extend our ability to understand contextual information by including semantic information in the scene. In addition, we plan to predict target goals and destinations as well as learn patterns of life where a sequence of targets can be learned. We also plan to integrate these algorithms into a robot planning framework and assess the robot's ability to navigate efficiently to the desired goal while avoiding obstacles. Finally, we plan to assess the human's reaction to the robot's navigation in daily environments.

References

[1] V. Karasev, A. Ayvaci, B. Heisele, and S. Soatto, "Intent-aware long-term prediction of pedestrian motion," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, May 2016, pp. 2543–2549.

[2] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[3] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras, "Human motion trajectory prediction: A survey," *CoRR*, vol. abs/1905.06113, 2019. [Online]. Available: http://arxiv.org/abs/1905.06113

[4] S. Pellegrini, A. Ess, and L. Van Gool, "Improving data association by joint modeling of pedestrian trajectories and groupings," 09 2010, pp. 452–465.

[5] L. Leal-Taix, M. Fenzi, A. Kuznetsova, B. Rosenhahn, and S. Savarese, "Learning an image-based motion context for multiple people tracking," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 3542–3549.