2019 IEEE 5th International Conference on Mechatronics System and Robots

Pedestrian Trajectory Prediction Using RNN Encoder-Decoder with Spatio-Temporal Attentions

Niraj Bhujel
School of Electrical and Electronics Engineering
Nanyang Technological University
Singapore, 639798
e-mail: bhuj0001@ntu.edu.sg

Wei-Yun Yau
Institute for InfoComm Research
A*STAR
Singapore 138577
e-mail: wyyau@i2r-a-star.edu.sg

Eam Khwang Teoh
School of Electrical and Electronics Engineering
Nanyang Technological University
Singapore, 639798
e-mail: eekteoh@ntu.edu.sg

***Abstract*—Pedestrian motion are inherently multi-modal in nature influenced by presence of other human and physical objects in the environment. Trajectory prediction models need to address both human-human and human-space interaction issues. In this work, we leverage both pedestrians information and scene information of the navigation environment for jointly predicting trajectories of the pedestrian. We introduce a new Recurrent Neural Network based sequence model with attention mechanisms that address both human-human and human-space interaction challenges. The encoder encodes all the pedestrian trajectories and create a social context. The scene information of navigation environment is extracted using CNN and serves as a physical context for the model. Our approach utilizes physical and social attention mechanism to find semantic alignments between encoder and decoder. The social attention mechanism allow the model to look into similar step of pedestrian trajectory. The physical attention mechanism tells the model where and what to focus on the scene. Experiment on several datasets shows that the proposed approach which combine social and physical attention performs better than when this information is utilized independently.**

Keywords-Deep Learning; LSTM; Attention Mechanism; Trajectory Prediction; Social Robot Navigation

I. INTRODUCTION

Autonomous robots operating in human environment are required to navigate safely and intelligently without causing discomfort to the other people in the space. Consequently, intelligent robots need the ability to predict the future motion of other pedestrian that allow them to plan the best path and collision avoidance strategy in advance. However, predicting the trajectory of pedestrian is challenging as there exists complex interaction with other dynamic and static humans/objects in the environment. The task of trajectory prediction should, therefore, address the influence of other pedestrians, i.e.human-human interaction and the effects of physical objects or obstacles present in the environment, i.e.human-space interaction. It is shown that with multi modal interaction model, more socially complaint robot navigation was achieved [22] [20].

Trajectory prediction approaches can be widely grouped into two major categories: model-based [12] [14] [13] and learning-based [23] [1] [24] [19]. While model based approach shows superior performance in static environments, learning-based approaches like Long Short-Term Memory(LSTM) based approach are found to perform better in the dynamic environments like crowds [24] [6] [19]. They are able to predict social behavior like slowing down, avoiding groups and following person [10].

An evaluation of various RNN based encoder-decoder net-works for trajectory prediction, [3], showed that Recurrent Encoder-Decoder was able to achieve better results compared to social force model [11] and social LSTM [1].In addition, LSTM frameworks allow for straightforward integration of motion information of other pedestrians and scene information over time into a single model [15] [19]. Moreover, recently, attention mechanism [2] are employed in RNN's architecture to allow the model to focus on the distinct aspects of the inputs and improving the quality of predicted trajectory [24].

In this work, we proposed an LSTM based encoder-decoder model that utilizes both pedestrian motion information and scene information to jointly predict the trajectory of all pedestrians in the scene. We extend the LSTM's framework, which are successfully applied in sequence predictions like handwriting [8] and speech generation [9], to solving pedestrian trajectory prediction problem. The challenge of LSTM to capture relevant dependencies between input and output is addressed by the attention mechanism. In contrast to existing approaches [1] [10] [19], our proposed model utilizes both social and physical information and hence is able to capture the multi-modal nature of the problem. Unlike previous approaches that use local grid around a pedestrian [1] [19] which do not consider pedestrian coming from opposite direction, our model takes into account all pedestrian in the scene.

978-1-7281-2223-6/19/$31.00 ©2019 IEEE

While[24] utilize one LSTM per pedestrian which is not scalable as crowd increase, we use only a single LSTM for all pedestrian in the scene for computation efficiency.

II. Related Work

Pedestrian trajectory prediction approaches can be widely categorized into model based and learning based methods. In model based approach, Social Force Model (SFM) [14] remain the most popular trajectory prediction algorithm so far. The pedestrian trajectory is estimated by measuring the virtual forces resulting from the interaction with other pedestrian and objects in the scene. The SFM has been extensively used in social-aware robot navigation [7] [8] [21] [25] and forecasting pedestrian trajectory [27]. SFM based pedestrian are suitable for modeling human-human interaction however heavily dependent on hand-crafted features and prior information about goal of the pedestrian which is not available on real-time application like robot navigation. A number of variations on the model was done with different model specification such as elliptical specification [13] and collision prediction [36].

By contrast, in learning based approach like Inverse Reinforcement Learning (IRL), the objective is to learn the cost function of robot path using expert demonstration and predict the trajectory with lowest cost. In, [29] [14], the authors extended the IRL based approach by learning a reward function that determine the joint probability distribution over trajectories of multiple people. Similarly, [4] used deep RL to learn motion planning policies instead of using imitation learning of human or other expert demonstrations. In addition, they show how to induce social norms into the model rather than learning them, however, features used for learning are mostly handcrafted.

Recently, data-driven methods using RNN has been used to replace traditional method like SFM and IRL because of it's suitability in sequence generation. [1] use LSTM to learn general human movement and predict their future trajectories in crowded space. [15] introduce a RNN Encoder-Decoder framework which uses variational auto-encoder (VAE) to learn static scene context and generate trajectories accordingly. In model like [28] [19], human-space and human-human interaction is considered by using *pedestrian* state, occupancy grid and CNN features which are fed into three channel hierarchical LSTM encoder.

The existing approaches predict single trajectory only and the interaction space among human is limited in local area defined by occupancy map [1] [19]. To address these limitation, [10] leverage the potential of Generative Adversarial Networks (GAN) taking into account of all humans and generate socially acceptable trajectories based on past trajectories only. Moreover, [24] proposed new GAN-based model able to capture the static scene context and past trajectories of all other pedestrians to address both human-space and human-space interaction.

Very few models [15] [24] [28] address the problem of multi-modal interactions by combining static scene and social aspects. However, it is still not clear how to utilize these information for creating socially-acceptable trajectory prediction. We rely on a theory from SFM [11] that pedestrian trajectories are influenced by interesting objects in the environment and position of other pedestrian in the environment. Our proposed model extract unique features of objects from the scene image and encode all other pedestrian positions into a single representation vector. In contrast to widely used personal zone of small radius around the pedestrian, we consider all pedestrian in the scene. We use visual attention to focus on the interesting objects in the environment and social attention to focus on social situation which enable our model to predict physically and socially aware steps during the prediction time.

III. Our Approach

A. Problem Definition

Trajectory prediction can be formally stated as predicting the future trajectories of an agent i given the past trajectories X_i of agent i and scene information. The agent's observed past trajectory X_i is given as:

$$X_i = \{(x_i^t, y_i^t) t = 1, \ldots, t_{obs}\}$$

The future ground truth trajectory of agent i is denoted as:

$$Y_i = \{(x_i^t, y_i^t) t = t_{obs} + 1, \ldots, t_{pred}\}$$

We denote the future predicted trajectory of agent i by $\hat{Y}_i$. The objective is to find the model:

$$\hat{Y}_i = F(X_i, H) for\ 1 \le i \le N$$

where, H is the hidden state of the model.

B. Maintaining the Integrity of the Specifications

The RNN encoder-decoder, proposed by [5] and [26], is a general method to learn the conditional distribution over a variable-length sequence, $y = (y_{1,\ldots}, y_{T'})$, conditioned on variable-length input sequence $x = (x_1, \ldots, x_T)$.

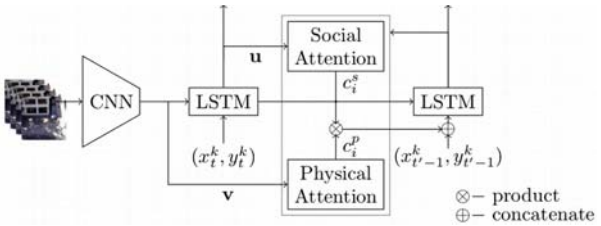

Figure 1. Overview of proposed deep LSTM prediction model

An overview of our proposed model is shown in Fig. 1. The overall model consists of four major steps; scene representation, encoding, attention and decoding. At first, the raw images are passed to CNN that represent the input images by the feature vectors v . The observed trajectories

of pedestrian k , $x=\left\{\left(x_1^k,y_1^k\right),\left(x_2^k,y_2^k\right),\ldots,\left(x_T^k,y_T^k\right)\right\}$, is then fed into the encoder one step at a time and produce a large fixed sized vector representation u .

$$h_t = LSTM([x_t^k, y_t^k], h_{t-1})$$

$$u = q(\{h_{1,\ldots}, h_T\}), t = 1,\ldots, T$$

where h_t is a hidden state at time t and q is non-linear activation function.

The decoder is trained to predict the next position $\left(x_{t'}^k,y_{t'}^k\right)$, given the previous hidden state $h_{t'-1}$,the previous predicted position $\left(x_{t'-1}^k,y_{t'-1}^k\right)$ and the context vector c t . The context vector is derived by multiplying the outputs from the social and physical attention module (more details in Section D). The decoding process is described by following equations:

$$h_{t'} = LSTM\left(h_{t'-1}, \left[x_{t'-1}^k, y_{t'-1}^k\right], c_t\right)$$

The network parameter is optimized using the loss function:

$$L = \frac{1}{T'}\sum_{t=1}^{T'} \|(x_t^k, y_t^k) - (\hat{x}_t^k, \hat{y}_t^k)\| + \lambda . \alpha, k = 1, 2, ..., N$$

where $\left(x_t^k,y_t^k\right)$ represent the ground truth position and $\left(\hat{x}_t^k,\hat{y}_t^k\right)$ is predicted position for the prediction horizon of T' , α is regularization term and λ is regularization factor.

C. Scene Representation

The relevant feature of the pedestrian environment is extracted using pre-trained CNN model for object detection. The CNN produces L features, each of D-dimensional vector. Thus the input scene can be represented by context vector corresponding to the features extracted at different image locations. The physical context vector is later used by physical attention module that direct the decoder to focus on important image locations. Like in [27], the features are extracted from the lower convolution layer as this allows the decoder to selectively focus on certain region of the image by selecting a subset of all the feature vectors.

D. Attention Mechanisms

As agent navigate in complex and dynamic environment, they focus on the small region of the scene like obstacles along the path, pavements, turnings or other pedestrian walking towards them. This is achieved with two different attention mechanisms: physical attention and social attention respectively.

Physical Attention: Physical attention aims to generate a context by attending to certain location of the input image. At each decoding step, t, the physical context vector c_i^p is computed as a weighted sum of the feature vectors $v_i, i = 1,\ldots,L$ For each location i in the image, a positive weight α_i is computed that can be understood as the probability that location i is the right place to focus for predicting the next position [2].

$$e_{ti} = \tanh\left(W_p h_{t-1}\right) + \tanh\left(U_p v_i\right)$$

$$\alpha_{ti}^p = \frac{\exp\left(e_{ti}\right)}{\sum_{k=1}^{L}\exp\left(e_{tk}\right)}, c_t^p = \sum_{i=1}^{L}\alpha_{ti}^p v_i$$

Social Attention: Social attention computes a social context vector c_i^s from the hidden state of the encoder $u=\left(h_1, h_2, \ldots, h_{t_{obs}}\right)$. Each hidden state h i is an encoder representation of the positions of all pedestrians thus encoding influence all pedestrian on current pedestrian.

Where Ws, Us are the network parameters and h_{t-1} is decoder previous hidden output.

$$e_{ti} = \tanh\left(W_s h_{t-1}\right) + \tanh\left(U_s v_i\right)$$

$$\alpha_{ti}^s = \frac{\exp\left(e_{ti}\right)}{\sum_{k=1}^{L}\exp\left(e_{tk}\right)}, c_t^s = \sum_{i=1}^{L}\alpha_{ti}^p v_i$$

E. Implementation Details

The physical features are extracted from the last convolution layer of VGG16 [25] that generate 14x14, 512 dimensional feature vector. We use a single-layered LSTM for encoder and decoder with 256 hidden state with dropout rate of 0.5. The model was trained using RMSProp for 200 epochs with learning rate of 0.001. All the parameter were uniformly initialized between -0.08 and 0.08. The model was implemented in Tensorflow. The training was done on single GEForce GTX 980 GPU.

IV. Expreiments

A. Dataset

We used two publicly available real-world datasets: ETH [18] and UCY [16]. The ETH datasets contains two scenes (ETH and Hotel) each with 750 pedestrians and annotated trajectories with position and velocity of each pedestrians. The UCY contains three scenes with 643 pedestrians from oblique view. In total, there are 5 sets of data consisting of 1393 pedestrians in crowded setting with challenging scenario like people crossing, overtaking, group forming/dispersing and collision avoidance.

B. Evaluation Metrics

The quantitative performance of all models is measured using two metrics:Average Displacement Error(ADE) and Final Displacement Error (FDE) which is a mean square error (MSE) or L2 error between the ground truth trajectory and all predicted trajectory as introduced by [18] and the

distance between the ground truth final position and predicted final position respectively. We did further study of the model with different configuration. We refer p_a and s_a as physical and social attention respectively.

C. Evaluation Methodology

Similar with [1], we use leave-one-out approach for training. The model was trained for four datasets and accuracy was tested on remaining set. This repeat until all datasets are trained and tested. The model is trained with 8 time steps (3.2 seconds) and accuracy was measured 12 time steps (4.8 seconds) of the predicted trajectory. The times-step length is similar with the widely used approach by [18] [1]. The quantitative performance is measured using ADE and FDE, whereas, the qualitative evaluation is done by visual inspection and comparison of the predicted trajectories among all models.

TABLE I: QUANTITATIVE RESULTS OVER FIVE DATASETS. THE ERROR METRICS AVERAGE DISPLACEMENT ERROR(ADE) AND FINAL DISPLACEMENT ERROR(FDE) IN NORMALIZED PIXELS IS USED FOR COMPARISON.

Metric	Dataset	Linear	RED	RED +pa	RED +sa	RED +sa+ pa
ADE	ETH	0.006	0.166	0.192	0.273	0.149
	Hotel	0.001	0.136	0.112	0.149	0.125
	Zara1	0.002	0.188	0.180	0.101	0.157
	Zara2	0.003	0.188	0.180	0.140	0.141
	Univ	0.002	0.233	0.366	0.208	0.148
	Average	0.002	0.174	0.208	0.174	**0.144**
FDE	ETH	0.012	0.301	0.332	0.466	0.293
	Hotel	0.002	0.271	0.335	0.301 1	0.264
	Zara1	0.003	0.308	0.330	0.200	0.304
	Zara2	0.004	0.390	0.327	0.325	0.298
	Uni	0.002	0.449	0.474	0.390	0.273
	Average	0.003	0.343	0.359	0.336	**0.286**

D. Quantative Results

The quantitative result is compared using two metrics ADE and FDE on ETH and UCY datasets. Table I shows the quantitative result of all models. In order to assess the performance of our model, we compared it against linear model that is implemented with off the shelf Kalman filter library. The linear model assumes that the velocity and acceleration of the each pedestrian are constant. The linear model have the lowest prediction error in terms of ADE and FDE among all model as expected. However do note that the linear model only serves as a lower bound for the motion predictions and capable of only modeling straight paths.

The RED model is a simple implementation of sequence model without attention mechanism which serves as a baseline for other variations. The first model RED+sa applies social attention and perform slightly better than the RED model. In RED+pa, physical attention is added and, as expected, themodel perform worse than the baseline RED model as it is not enough to truly understand the social behavior of the pedestrians from physical contexts only. The major improvement in model performance comes when we combine physical and social attention together in RED+sa+pa. The RED+sa+pa is able to learn the social and physical contexts thereby outperforming the baseline models. From the experimental result, it can be concluded that combining physical and social attention is advantageous and allow for more robust trajectory predictions in crowded environment.

E. Qualitative Results

In this section we further investigate our model by visualizing the predicted trajectories (see Figure 2) and the attention component learned by the model (see Figure 3). We refer to each pedestrian in the scene by the first letter of the color in the figures (e.g. person G(green) and so on) It is clear that the linear model is able to predict the straight path only while the RED model can predict complex trajectories. The RED model is however, still prone to false prediction that sometimes goes off the pedestrian path. The prediction error was improved by using physical and social attention as shown in Figure 2 (d). In order to visualize the attention weights, we up-sample it by a factor of 2 4 = 16 and apply Gaussian filter (see Figure 3). The input image is resized to 224x224 before feeding into the network. The attention weights of the region is highlighted with the green color. The model learn to focus on different multiple region over time and take appropriate action accordingly.

V. CONCLUSION

In this work, we tackle the problem of trajectory prediction by considering both human-human and human-space interaction. We introduce a new RNN encoder decoder based architecture with attention mechanisms utilizing social and physical information. A key advantage of our proposed approach is that we used one LSTM for predicting trajectories of multi-pedestrian which reduced computational complexity significantly. This is important in crowded environment and especially suitable when density changes quickly. The model doesn't require destination of pedestrian, which makes it suitable for real-world application. From the experiment result, RNN model with both physical and social attention mechanisms perform better than when this information is used separately.

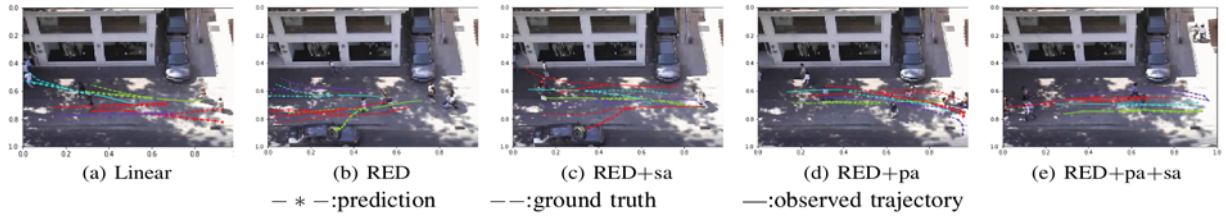

Figure 2. Illustration of predicted trajectories by different models.

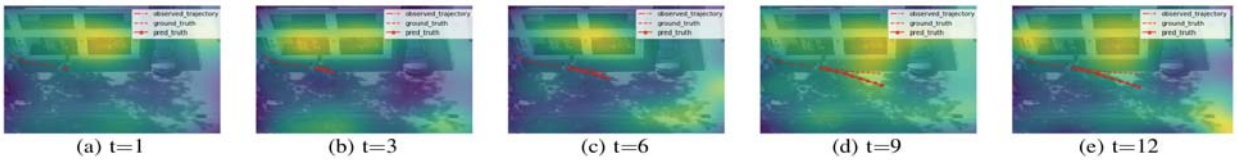

Figure 3. Distribution of physical attention over time. At each time step, the attention changes different parts of the image.

ACKNOWLEDGEMENT

This research is partially supported by SERC grant No. 162 25 00036 from the National Robotics Programme (NRP), Singapore.

REFERENCES

[1] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *CVPR*, 2016.

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. ArXiv:1409.0473, 2014.

[3] S. Becker, R. Hug, W. Hübner, and M. Arens. An evaluation of trajectory prediction approaches and notes on the trajnet benchmark. arXiv:1805.07663, 2018.

[4] Y. F. Chen, M. Everett, M. Liu, and J. P. How. Socially aware motion planning with deep reinforcement learning. In *IROS*. IEEE, 2017.

[5] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv:1406.1078, 2014.

[6] T. Fernando, S. Denman, S. Sridharan, and C. Fookes. Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection. In *Neural networks*, 108:466–478, 2018.

[7] G. Ferrer and A. Sanfeliu. Proactive kinodynamic planning using the extended social force model and human motion prediction in urban environments. In *IROS*. IEEE, 2014.

[8] A. Graves. Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850, 2013.

[9] A. Graves, A. R. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *ASSP*. IEEE, 2013.

[10] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *CVPR*, 2018.

[11] D. Helbing and A. Johansson. Pedestrian, crowd and evacuation dynamics. In *Encyclopedia of Complexity and Systems Science*, pages 6476–6495. Springer, 2009.

[12] D. Helbing and P. Molnar. Social force model for pedestrian dynamics. In *Physical Review* E, 51(5):4282–4286, 1995.

[13] H. Kretzschmar, M. Spies, C. Sprunk, and W. Burgard. Socially compliant mobile robot navigation via inverse reinforcement learning. In *IJRS*, 2016.

[14] M. Kuderer, H. Kretzschmar, C. Sprunk, and W. Burgard. Feature-based prediction of trajectories for socially compliant navigation. In Robotics: Science and Systems. Citeseer, 2012.

[15] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. Torr, and M. Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *CVPR*, 2017.

[16] A. Lerner, Y. Chrysanthou, and D. Lischinski. Crowds by example. In Computer Graphics Forum, volume 26, pages 655–664, 2007.

[17] P. Patompak, S. Jeong, N. Y. Chong, and I. Nilkhamhang. Mobile robot navigation for human-robot social interaction. In International Conference on Control, Automation and Systems, Oct 2016.

[18] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV, IEE*E, 2009.

[19] M. Pfeiffer, G. Paolo, H. Sommer, J. Nieto, R. Siegwart, and C. Cadena. A data-driven model for interaction-aware pedestrian motion prediction in object cluttered environments. CoRR, 2017.

[20] M. Pfeiffer, M. Schaeuble, J. Nieto, R. Siegwart, and C. Cadena. From perception to decision: A data-driven approach to end-to-end motion planning for autonomous ground robots. In *ICRA*. IEEE, 2017.

[21] E. Repiso, G. Ferrer, and A. Sanfeliu. On-line adaptive side-by-side human robot companion in dynamic urban environments. In *IROS*. IEEE, 2017.

[22] J. Rios-Martinez, A. Spalanzani, and C. Laugier. From proxemics theory to socially-aware navigation: A survey. International Journal of Social Robotics, 7(2):137–153, 2015.

[23] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *ECCV*, 2016.

[24] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, and S. Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. arXiv:1806.01482, 2018.

[25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. ArXiv:1409.1556, 2014.

[26] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112, 2014.

[27] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[28] H. Xue, D. Q. Huynh, and M. Reynolds. Ss-lstm: A hierarchical lstm model for pedestrian trajectory prediction. In *WCACV*, IEEE 2018.

[29] B. D. Ziebart, N. Ratliff, G. Gallagher, C. Mertz, K. Peterson, J. A. Bagnell, M. Hebert, A. K. Dey, and S. Srinivasa. Planning-based prediction for pedestrians. In *IROS*. IEEE, 2009.