

Noticing Motion Patterns: A Temporal CNN with a Novel Convolution Operator for Human Trajectory Prediction

Dapeng Zhao¹ and Jean Oh¹

Abstract—As more and more robots are envisioned to cooperate with humans sharing the same space, it is desired for robots to be able to predict others’ trajectories to navigate in a safe and self-explanatory way. In this paper, we propose a Convolutional Neural Network-based approach to learn, detect, and extract patterns in sequential trajectory data, known here as Social Pattern Extraction Convolution (Social-PEC). A set of experiments carried out on the human trajectory prediction problem shows that our model performs comparably to the state of the art and outperforms in some cases. More importantly, the proposed approach unveils the obscurity in the previous use of a pooling layer, presenting a way to intuitively explain the decision-making process.

Index Terms—Intention Recognition, Social HRI, Deep Learning Methods, Motion and path planning

I. INTRODUCTION

WHETHER an agent is co-working with other agents or is navigating in a crowd, it is critical for the agent to be able to understand and predict the motions of other agents in the same environment to ensure safe interaction and efficient performance. In this work, we particularly focus on the problem of pedestrian trajectory prediction in crowded environments. Predicting pedestrian behavior is challenging because pedestrians’ future trajectories can be affected not only by the physical properties that we have well-established models to explain, such as energy or momentum, but also by the pedestrians’ hidden objectives and subtle social norms in crowd interactions.

The majority of existing work in pedestrian trajectory prediction generally follows the encoder-decoder model where the past trajectories are summarized to capture the context of the crowd movements. In this paper, we propose an approach, known here as Social Pattern Extraction Convolution (Social-PEC), where trajectories are eventually represented as a combination of *motion patterns*. Whereas existing approaches take the raw trajectories to encode the social context, the proposed idea on motion patterns is to generalize the observed trajectory as illustrated in Fig. 1. We design our model to understand and

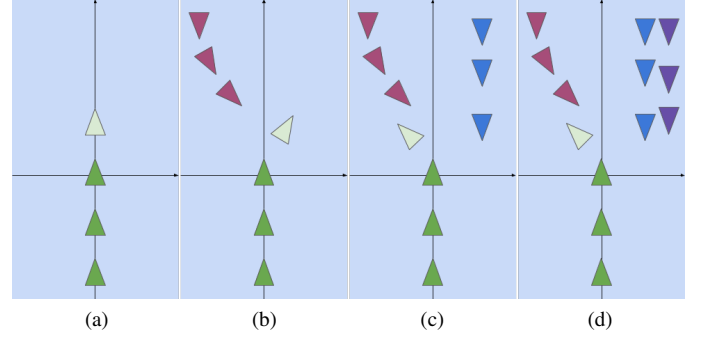


Fig. 1. Four sample scenarios of pedestrian interaction. The green triangles and the light green triangle represent the target pedestrian’s history trajectories and expected one-step future location; triangles in other colors, the history trajectories of other pedestrians; and the sharp tips, the trajectory directions. In (a), no motion patterns from others are present; therefore, the target is expected to proceed linearly. In (b), the motion pattern of “somebody approaching me from my front left” is present, which impacts our target to walk to the right in order to avoid a possible collision. In (c) and (d), another motion pattern of “somebody on my right is about to pass by me” is present; therefore, our target pedestrian walks towards left.

learn the various motion patterns of pedestrians in crowds, force the model to “notice” them during training and predict upon them. We build our sequence encoder using the idea of Temporal Convolutional Neural Networks (CNN) [1] and propose a new convolution operator (defined in III-B) that enables our model to actually detect, learn, and extract motion patterns from the observed trajectories. Using a different convolution operator is not a new idea: in [2], the conventional correlation-based convolution operator has been modified to successfully achieve satisfying performance on the MNIST dataset, showcasing the applicability of the generalization of convolution operation in CNNs.

Our model, Social-PEC, achieves comparable results with the state-of-the-art methods on public datasets in terms of standard evaluation metrics based on the displacement errors. Additionally, the use of motion patterns unveils the obscurity in social pooling, and makes the decision making process more transparent, intuitive, and explainable.

II. RELATED WORK

Several technical approaches have been used to tackle the pedestrian trajectory prediction problem. In early works, algorithm designers tried to assert domain knowledge about social interactions in crowds to algorithms explicitly, e.g., Social

Manuscript received: July 1, 2020; Revised November 12, 2020; Accepted December 7, 2020.

This paper was recommended for publication by Editor Tamim Asfour upon evaluation of the Associate Editor and Reviewers’ comments.

This work was supported in part by the Air Force Office of Scientific Research and U.S. Army Ground Vehicle Systems Center.

¹Dapeng Zhao and Jean Oh are affiliated with the Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA 15213 {dapengz, hyaejino}@andrew.cmu.edu

Digital Object Identifier (DOI): see top of this page.

Force [3] and Interactive Gaussian Process (IGP) [4], [5]. However, some researchers later suggested that hand-crafting models and rules have various limitations. These findings have led to those approaches that would allow the machines to learn directly from data [6], [7], [8], resulting in significantly improved performances in general. We will mainly discuss the latter data-driven type of techniques in this section.

For modeling sequential data, Recurrent Neural Network (RNN) and its variations such as Long Short-Term Memory network (LSTM) have been the popular choice, e.g., Social-LSTM [7]; however, the benefit or necessity of using RNNs for pedestrian trajectory prediction is debatable. RNNs tend to gradually forget information from the past, hence the idea of LSTM was proposed as a remedy to selectively forget/remember. An LSTM is a reasonable choice in many problem domains where the sequences can be arbitrarily and extensively long, e.g., in the text-related tasks, a sequence could be a lengthy article where important contextual information can appear anywhere in the text. By contrast, in the pedestrian trajectory prediction problem, except for some rare extreme cases, the models do not need excessively long sequences as their inputs, as it is generally enough to observe how people have moved in the last 5 seconds, i.e., the observations from far too past are no longer meaningful to the current interaction, if not misleading at times. Moreover, RNNs/LSTMs have other drawbacks [9] including vanishing/exploding gradients, unstable and expensive training, and inefficient parameters. Remarkable efforts were also made by Bai et al. [9] and Becker et al. [10] to empirically evaluate RNNs for sequential data learning.

After each sequence is modeled and encoded, information needs to be aggregated together, for which a pooling layer has been a popular choice. A pooling layer is widely used in CNNs for the image processing tasks, typically after spatial convolutional layers [11]. In these CNNs, the vectors being pooled are quite literally the correlation between the kernels and the signals, and a pooling operation is to extract the strongest signal in a local region. However, the hidden space of RNNs/LSTMs is not well understood, and it is difficult, if not impossible, to understand its semantic meanings in a physical space. That said, it is yet to be justified to use a pooling layer after RNN/LSTM's modelling sequences for information aggregation. On this note, Mohamed et al. [12] reached a consensus with us.

Another popular design choice is to use the graph representation. The combination of a graph with RNNs/LSTMs [13], [14], and the combination of a graph with CNNs (Graph CNN) [12] have both been proposed. In these approaches, however, the size of a graph is generally dependent on the number of pedestrians in the scene, so it can face the scalability challenges as the number of pedestrians grows significantly in crowded scenes.

An extensive and comprehensive survey article for human motion prediction is done by Rudenko et al. [15], where interested readers can find additional relevant works.

In contrast to the existing approaches, the proposed Social Pattern Extraction Convolution (Social-PEC) model avoids the issues pointed out above in this section. Firstly, it takes

a fixed length of history trajectory as inputs and does not rely on RNNs to encode trajectory, thus avoids the RNN training issue. Secondly, the trajectory encoder is based on "motion patterns" that are intuitively reasonable and can be easily visualized, thus avoids the obscurity during "social pooling" which is to aggregate the encoded embedding of the neighboring pedestrians' trajectories.

III. TRAJECTORY PREDICTION WITH PEC

Our approach builds on an assumption that, when navigating in a crowd, humans react to an abstract representation of a scene, e.g., at the level of motion patterns that they have seen frequently in their past experiences. For instance, whether they are 2 or 3 people, whether they walk slightly faster or slower, whether they are a few centimeters to the left or right, as long as they approach from the same general direction at roughly the same distance with similar speeds, we probably will react very similarly as shown in Fig.1. Although trajectory data are mostly stored as sequences of location coordinates, humans do not react to precise location coordinates; instead, we react to general *motion patterns*.

In this paper, we define a "pattern" to be a segment of data. Specifically, a "motion pattern" refers to a short segment of trajectory that can be frequently seen in real trajectory data. A motion pattern is represented as a sequence of location coordinates, similar to a short trajectory representation.

Logistically, our strategy is to only predict one-step future locations, and use the predicted locations as if they are the new observations to further predict. When predicting a one-step future, we predict for each pedestrian at a time. The pedestrian we predict for is referred to as a "target pedestrian", while all others are "context pedestrians", and all trajectories are transformed from the world coordinates to the target pedestrian's egocentric coordinates where the target pedestrian is at the origin, facing the positive direction of x-axis.

In this section, we first clarify problem setup and notation; then introduce our model's key component, Pattern Extraction Convolution (PEC); next, we introduce the actual trajectory predictor; finally, we explain training and inference.

A. Representation and Notation

Suppose that there are M pedestrians in a scene. Given all of their observed history trajectories in the world coordinates, the goal is to predict the future trajectories of all pedestrians.

A trajectory is a series of states timestamped at a constant interval. State s is defined as a 2-dimensional location coordinates, s.t., $s = (x, y)$. We note that the definition of a state can be extended to include additional information such as orientation or personality in the future.

In this paper, we use a finite length T_f of timesteps. The start and end time for history observations are 1 and T_h , and that of future prediction are $T_h + 1$ and T_f .

For pedestrian $m \in \{1, \dots, M\}$, the trajectory of m observed from timestep 1 to timestep T_h , denoted by $\phi_{1:T_h}^m$, is composed of a sequence of (x, y) coordinates as follows:

$$\phi_{1:T_h}^m = \{s_t^m | t \in \{1, \dots, T_h\}\} \quad (1)$$

where $\mathbf{s}_t^m = (x_t^m, y_t^m)$ is pedestrian m 's state at time t .

Let Φ denote the set of trajectories, s.t., $\Phi[m] = \phi^m, \forall m \in \{1, \dots, M\}$; the dimension of Φ is thus $(M, T_f, 2)$.

B. Pattern Extraction Convolution (PEC)

Pattern Extraction Convolution (PEC) is a mechanism that, in an intuitive sense, detects and recognizes patterns from data, while effectively projects trajectories from the x-y coordinate space to a new space that is defined in terms of similarities between patterns and the trajectory.

Let \mathbf{P} denote the set of motion patterns, and let there be N patterns in total. The j -th motion pattern, $\forall j \in \{1, \dots, N\}$, is defined as the following:

$$\mathbf{P}[j] = \{\mathbf{s}_t^j | t \in \{1, \dots, L\}\} \quad (2)$$

where L is the pattern length, $\mathbf{s}_t^j = (x_t^j, y_t^j)$ is the state. The dimension of \mathbf{P} is thus $(N, L, 2)$.

We define PEC as an encoder that translates raw trajectory ϕ to abstract trajectory ψ in terms of motion patterns \mathbf{P} as follows:

$$\psi = \text{PEC}(\phi; \mathbf{P}) \quad (3)$$

At each timestep t , for each motion pattern j , the PEC operation is defined as:

$$\begin{aligned} \psi[t, j] &= \text{PEC}(\phi; \mathbf{P})[t, j] \\ &= \lambda[j] \cdot \log\left(\sum_k \left\| \phi[t-L+k, :] - \mathbf{P}[j, k, :] \right\|_2\right) + \mathbf{b}[j] \end{aligned} \quad (4)$$

where λ is a scaling coefficient, and \mathbf{b} , biases. The dimension of the resulting encoded trajectory ψ is $(T_h - L + 1, N)$ where each entry of $\phi[t, j]$ indicates the similarity between the corresponding segment of the trajectory and pattern $\mathbf{P}[j]$. This operation is further demonstrated in Fig.2.

Because the output ψ of the PEC operation will later interact with an activation function such as \tanh , it is necessary to bring in the log function and the extra scaling coefficients λ . Specifically, the log function helps re-range the values of the L2-distances from $[0, +\infty)$ back to $(-\infty, +\infty)$. Scaling coefficients λ helps scale the output properly to better interact with the non-linearity of the activation function. These two practices were not necessary for the conventional convolution operator because the dot-product operation naturally ranges $(-\infty, +\infty)$ where the magnitude of a kernel matrix adds already an extra degree of freedom to help scale the response.

It is worth noting that the PEC operator defined in (4) is different from the conventional convolution (CONV) operator [2], [16] which is commonly used in CNNs for image-related tasks. The main differences are (1) PEC is based on the L2-difference to measure the physical distance whereas CONV is based on dot-product; (2) the conventional operator ignores the physical meaning of the channels and simply sums up the outputs from different channels. The necessity for using the PEC operator for trajectory encoding is further illustrated in Fig.3.

The set of motion pattern, \mathbf{P} , is learned from data, trained using the prediction loss through back propagation. More implementation details and illustration are presented in Sections IV-B and IV-E.

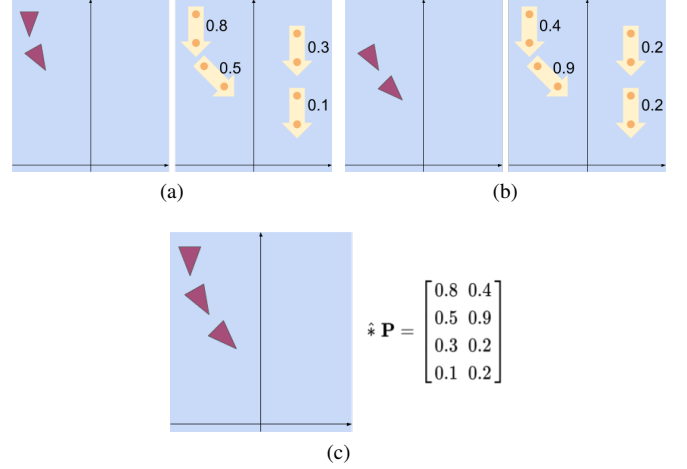


Fig. 2. An example showing how raw trajectories can be projected to the motion pattern space. Red triangles represent the input trajectory. The full trajectory ϕ is shown in (c) which is of shape $(3, 2)$, for $T = 3$. Yellow arrows represent the set of motion patterns of shape $(4, 2, 2)$, for the number of patterns $N = 4$, the length of pattern $L = 2$, and $\mathbf{s} = (x, y)$ is 2-dimensional. In (a), the similarities between the first segment of the trajectory and each motion pattern are found and marked next to the arrow; in (b), the second segment, similarly. The full operation on the trajectory level is shown in (c), where the resulting matrix is ψ in the shape of $(2, 4)$ for $T_h - L + 1 = 2$ and $N = 4$. The columns of this matrix ϕ are similarity scores from (a) and (b), respectively.

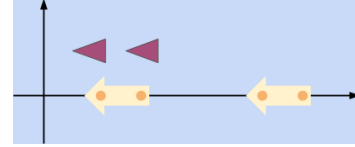


Fig. 3. An example showing the incompetency of the conventional convolutional operator for trajectories. Suppose red trajectory $\phi = \{(10, 1), (20, 1)\}$, and two patterns p_0 and p_1 are $\{(10, 0), (20, 0)\}$ and $\{(50, 0), (60, 0)\}$. Ideally, the magnitude of the convolution response (or output) is determined by the similarity only, i.e., the larger the more similar, but if CONV is used the magnitude of response would also be influenced by the magnitude of the input signals or kernels. In the example, the similarity between ϕ and p_0 should be larger than p_1 , but $\text{CONV}(\phi, p_0) = 500 < 1700 = \text{CONV}(\phi, p_1)$.

C. Human Trajectory Predictor

Our strategy is to predict one-step future locations first, and then use the predicted locations as if they are new observations to further predict. At each timestep, every pedestrian is treated as target pedestrian by turns.

For each target pedestrian $m \in \{1, \dots, M\}$, the location coordinates of all trajectories are first transformed from the world coordinates to the new coordinates where pedestrian m is at the origin, oriented towards the positive direction of the x-axis, as in the `Convert` function in Algorithm 1. Next, Location Predictor takes all of these trajectories as inputs, and outputs m 's next location, which is modeled as bivariate Gaussian distribution [7]. Finally, this location is converted back to the world coordinates and appended to Φ for location prediction for the next timestep.

Given target pedestrian m , let $-m$ denote the rest of the pedestrians except m , referred to as Context. First, $\Phi'[m]$ and $\Phi'[-m]$ are encoded respectively by two different encoder networks. The reason for having two different networks is that

Algorithm 1: Trajectory Predictor

```

1  $\Phi \leftarrow \{\phi^m | m \in [1 : M]\}$ 
2 Function TrajPredictor( $\Phi$ ):
3   for  $t = T_h + 1, \dots, T_f$  do
4     for  $m = 1, \dots, M$  do
5        $\Phi' \leftarrow \text{Convert}(\Phi[:, t - T_h : t], m)$ 
6        $\mu', \Sigma' \leftarrow \text{LocPredictor}(\Phi'[m], \Phi'[-m])$ 
7        $s' \sim \mathcal{N}(\mu', \Sigma')$ 
8        $s^m \leftarrow \text{ConvertBack}(s')$ 
9        $\Phi[m, t] \leftarrow s^m$ 
10  return  $\Phi[:, T_h + 1 : T_f]$ 

```

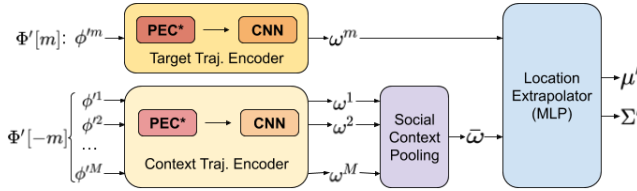


Fig. 4. Location Predictor that predicts one-step location for pedestrian m . Φ' is the set of trajectories already transformed to target pedestrian's coordinates. ω is the trajectory embedding, explained in Equation (5).

motion pattern set and context trajectories are expected to be different for Context and Target. Trajectory Encoder works on one trajectory at a time. The Target trajectory embedding ω can be written as follows:

$$\begin{aligned} \omega &= \text{CNN}(\sigma(\psi)) \\ &= \text{CNN}(\sigma(\text{PEC}(\phi; \mathbf{P}))) \end{aligned} \quad (5)$$

where ϕ is the raw trajectory in location coordinates space; ψ , the trajectory encoded by PEC; and σ , the activation function. By applying CNN to the encoded trajectory ψ , combinations of basic motion patterns are further extracted by the CNN on higher levels as more sophisticated patterns.

Now that the model is ready to predict future trajectory for target pedestrian, observations of all other pedestrians should be aggregated to provide the social context. Context Pooling layer is applied to compute context trajectory $\bar{\omega}$ as follows:

$$\bar{\omega}[t, k] = \max_{m \in [-m]} (\omega^m[t, k]) \quad (6)$$

where $\bar{\omega}$ has the same shape as the target's encoded trajectory ω^m , t and k are time and pattern indices.

To make predictions, both encoded target trajectory ω^m and social context $\bar{\omega}$ are fed into Multilayer Perceptron (MLP) in the end:

$$x, y, a, b, c = \text{MLP}(\omega^m, \bar{\omega}). \quad (7)$$

The μ' and Σ' of location's Gaussian distribution are then constructed from the raw MLP outputs:

$$\begin{aligned} \mu' &= [x, y] \\ \sigma'_{xx}, \sigma'_{yy} &= \exp(a), \exp(b) \\ \sigma'_{xy} &= \sigma'_{xx} * \sigma'_{yy} * \tanh(c) \\ \Sigma' &= \begin{bmatrix} \sigma'_{xx} & \sigma'_{xy} \\ \sigma'_{xy} & \sigma'_{yy} \end{bmatrix} \end{aligned} \quad (8)$$

where superscript $'$ indicates that variable is in the target-centered coordinates instead of the world coordinates. Construction in (8) is necessary because the value range of raw MLP outputs, $[-\infty, \infty]$, does not satisfy the constraints of the covariance matrix.

D. Training and Inferencing

During training, the future prediction length is set to be 1, s.t., $T_f = T_h + 1$, because all parameters that need to be trained are all within Location Predictor.

Parameters are learned by minimizing the following negative log-likelihood loss:

$$L = - \sum_{m=1}^M \log(P(\Phi'[m, T_h + 1] | \mu'^m, \Sigma'^m)) \quad (9)$$

During inferencing, the future prediction can have an arbitrary length, and the prediction output is $\Phi[:, T_h + 1 : T_f]$ as stated in Algorithm 1.

IV. EVALUATION AND DISCUSSION

A. Datasets and Metrics

Our model is evaluated on two datasets: [18] and [19]. They contain 5 crowd sets in different scenes with a total number of 1,536 pedestrians exhibiting complex interactions such as walking together, groups crossing each other, joint collision avoidance and nonlinear trajectories.

As for metrics, like [7], [8], [17], [14], [12], we use Average/Final Displacement Error (ADE/FDE) [12], which have been conventionally used for this problem.

In order to make full use of the data for evaluation and also to evaluate how well models generalize to unseen datasets, we use the leave-one-out approach where a model is trained and validated on 4 datasets and tested on the remaining set. To ensure a fair comparison, we use identical dataset step and train/validation/test split which are also used in S-SLSTM[7], S-GAN[8] and S-STGCNN[12].

The data used in our work are annotated every 0.4 seconds. Observation length is set to be 8 timesteps (3.2 sec) and future prediction length is set to be 12 timesteps (4.8 sec).

B. Implementation Details

As shown in Fig.4, our model mainly contains 3 modules, Context Trajectory Encoder, Target Trajectory Encoder, and Location Extrapolator.

Context Trajectory Encoder consists of a Pattern Extraction Convolution (PEC) layer and a conventional convolution (CONV) layer, each followed by activation function \tanh , with a max pooling layer in between. For the PEC and CONV, the number of kernels are 100 and 160, the kernel lengths are 2 and 2; the pooling stride is 2. If the number of input channels is 2 (x-y coordinates) and the data temporal length is 8 (length of observations), s.t. an input is in the shape of (2,8), the resulting output's shape will be (160,3).

Target Trajectory Encoder is very similar to Context Trajectory Encoder, except that the number of kernels are only 50 and 80, because there are much less varieties among target

TABLE I

ADE/FDE IN METERS FOR DIFFERENT METHODS, THE LOWER THE BETTER. LINEAR EXTRAPOLATION IS INCLUDED AS BASELINE AND ITS OUTPUT IS DETERMINISTIC, WHEREAS ALL OTHER MODELS PREDICT 20 TRAJECTORIES AT ONCE, AND ONLY THE BEST ONE IS COUNTED FOR EVALUATION.

Model	ETH	Hotel	Univ.	Zara1	Zara2	Ave.
Linear	1.33 / 2.94	0.39 / 0.72	0.82 / 1.59	0.62 / 1.21	0.77 / 1.48	0.79 / 1.59
S-LSTM[7]	1.09 / 2.35	0.79 / 1.76	0.67 / 1.40	0.47 / 1.00	0.56 / 1.17	0.72 / 1.54
SGAN(20VP20)[8]	0.87 / 1.62	0.67 / 1.37	0.76 / 1.52	0.35 / 0.68	0.42 / 0.84	0.61 / 1.21
STSGN[17]	0.75 / 1.63	0.63 / 1.01	0.48 / 1.08	0.30 / 0.65	0.26 / 0.57	0.48 / 0.99
S-BiGAT[14]	0.69 / 1.29	0.49 / 1.01	0.55 / 1.32	0.30 / 0.62	0.36 / 0.75	0.48 / 1.00
S-STGCNN[12]	0.64 / 1.11	0.49 / 0.85	0.44 / 0.79	0.34 / 0.53	0.30 / 0.48	0.44 / 0.75
Social-PEC	0.61 / 1.11	0.31 / 0.52	0.47 / 0.82	0.43 / 0.77	0.35 / 0.60	0.43 / 0.76

trajectories because the irrelevant variance have been removed by transforming the coordinate system with respect to target trajectory in Convert of Algorithm 1.

Location Extrapolator are 4 fully-connected layers, followed by leaky Re-LU activation. The width of the layers are 300, 120, 80, and 5.

For training, the batch size is 64, and it trains for 150 epochs using the Adam Optimizer [20] with the learning rate set as 0.001. The model is trained on GeForce RTX 2080 Ti.

C. Quantitative Results

As shown in Table 1, our model outperforms almost all models and performs comparably well with the current state-of-the-art, Social-STGCNN [12]. Other works are mostly RNN-based, while ours and Social-STGCNN are CNN-based. Our results suggest that CNNs might be a better option for modeling some types of sequences, e.g., pedestrian trajectories; more in-depth discussions related to the comparison between CNN and RNN can also be found in [10] and [9].

D. Qualitative Analysis

Some sample results are shown in Fig.5. Typically, linear trajectories are trivial to predict; however, our results seem to support that the proposed model also performs well for some non-linear trajectories, especially in more crowded scenes, e.g., red in (a), purple and orange in (b), and brown in (f). The success here might indicate that the proposed model is able to make use of social context effectively to make more accurate and more reasonable predictions.

Some of the predictions deviate from the true future trajectories significantly, e.g., brown in (a), red in (e), and pink in (f). These predictions, however, still appear reasonable, that is, based on the observed trajectories, the prediction may appear arguably more reasonable than the true future. Such “mispredictions” are inevitable to some extent as some of the observed history trajectories might not carry enough information to allow anyone to accurately predict their future.

Sometimes, the proposed is able to recognize the groups in crowd and predict accordingly, although we did not explicitly design the model to incorporate social group awareness [21]. In (f), the history observations of orange and blue are highly similar, thus their future trajectories are predicted to be very similar too, even the way how they deviate from ground truth are also similar. Comparatively, also in (f), the history

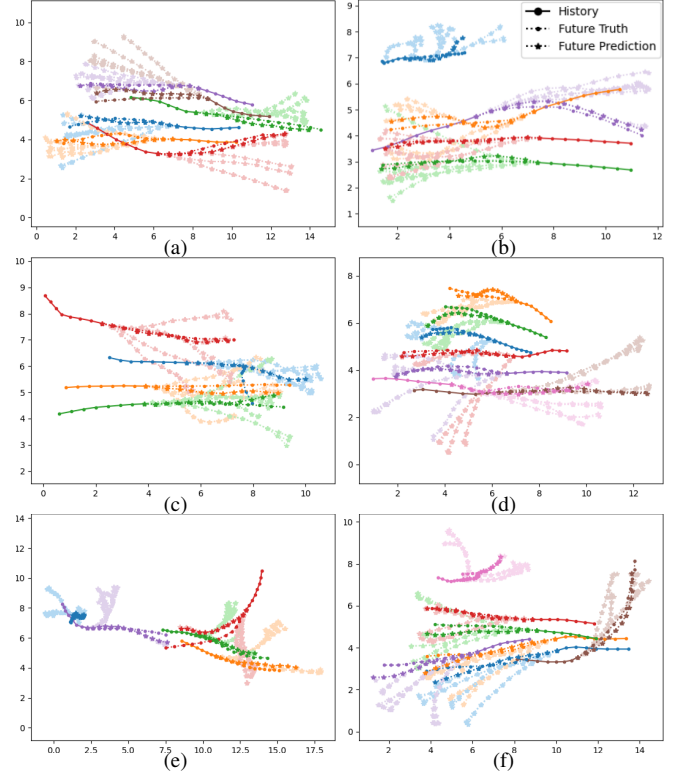


Fig. 5. Some samples of our model’s prediction. Our model predicts 20 possible future trajectories for each pedestrian, of which the one with the smallest ADE is highlighted. This figure is best viewed in color.

observations of red and green did not demonstrate enough similarity, thus in the model’s prediction they do not walk together any more.

In some cases, the model successfully shows appropriate precaution for collision avoidance. In (e), the interaction between red and green is notable. Red took a different path that is farther away from green and both were predicted to move slower than ground truth, a plausible explanation is that the proposed model was trying to avoid the two pedestrians colliding with each other. In the same scene, the orange is predicted to move faster than ground truth, because its front space appears to be clear enough to allow faster speed.

E. Motion Pattern Illustration

The weights for the PEC layer in Trajectory Encoder are motion patterns. Fig.6 presents two sets of patterns trained

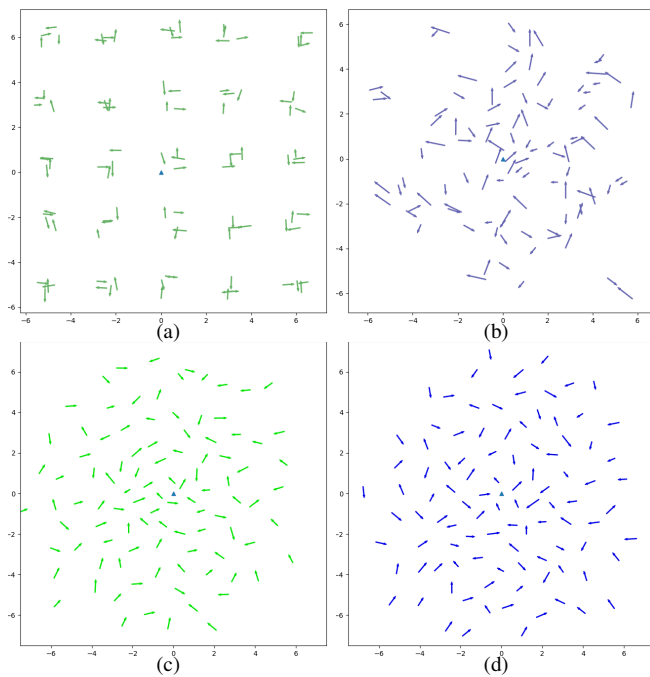


Fig. 6. Two sets of initialization and learned motion patterns for Context Trajectory Encoder. (a) and (b) are initialization, while (c) and (d) are their corresponding learned patterns after training. The number of patterns is 100; length of patterns is 2. The coordinates are with respect to the target pedestrian marked by a blue triangle at (0,0).

from two very different initialization, which demonstrates that the training is robust enough to allow different initialization to converge to similar patterns. It is noted that the learned patterns are not identical though similar; however, no significant performance difference has been observed in terms of ADE/FDE.

Fig.7 illustrates the core idea of “Noticing Motion Patterns” with an example scenario, where observed history trajectory data are matched to learned motion patterns at each timestep. This is done by the first layers of Trajectory Encoders, Pattern Extraction Convolution (PEC) layers.

Under the scheme of “motion pattern”, social pooling finally is no longer obscure. We are not pooling in any unknown latent space, but pooling in well-understood pattern space. The physical meaning of each entry in tensor ω is the similarity indicating how much of that particular motion pattern is present in the current scene. The larger the entry value is, the more similar the raw trajectory is to the motion pattern. Thus, stronger motion patterns should have a bigger impact on the target pedestrian’s decision making. By only considering the prominent presence for each motion pattern, the model can already be well-informed about its social context, of which Fig.1(c)(d) are good examples.

V. CONCLUSION

In this paper, we propose a CNN-based model for human pedestrian trajectory prediction with the idea of motion patterns. The main contributions of this work include:

- we present Pattern Extraction Convolution (PEC) as an intuitive and explainable mechanism to learn, detect,

and extract patterns from data, which is used to encode trajectories in this work;

- we further apply PEC to human trajectory prediction problem as the model of Social-PEC, and achieve comparable performance with the current state-of-the-art;
- the use of PEC avoids the obscurity in information aggregation (pooling layer) that was present in the previous work; and
- this study further challenges the community to re-examine the use of RNN in sequential data learning tasks.

In the future, for the problem of human motion prediction, we will try to incorporate physical environmental constraints into our model. Since the proposed idea on the PEC encoder is more general than the scope of trajectory prediction, we plan to apply PEC to other problem domains and further explore its potential.

ACKNOWLEDGEMENT

This work is in part supported by the U.S. Air Force Office of Scientific Research under award number FA2386-17-1-4660 and U.S. Army Ground Vehicle Systems Center.

REFERENCES

- [1] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition,” *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.
- [2] K. Ghiasi-Shirazi, “Generalizing the convolution operator in convolutional neural networks,” *Neural Processing Letters*, vol. 50, no. 3, pp. 2627–2646, 2019.
- [3] D. Helbing and P. Molnar, “Social force model for pedestrian dynamics,” *Physical review E*, vol. 51, no. 5, pp. 4282–4286, 1995.
- [4] P. Trautman and A. Krause, “Unfreezing the robot: Navigation in dense, interacting crowds,” in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 797–803.
- [5] P. Trautman, J. Ma, R. M. Murray, and A. Krause, “Robot navigation in dense human crowds: the case for cooperation,” in *2013 IEEE International Conference on Robotics and Automation*, 2013, pp. 2153–2160.
- [6] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert, “Activity forecasting,” in *European Conference on Computer Vision*. Springer, 2012, pp. 201–214.
- [7] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, “Social lstm: Human trajectory prediction in crowded spaces,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 961–971.
- [8] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, “Social gan: Socially acceptable trajectories with generative adversarial networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2255–2264.
- [9] S. Bai, J. Z. Kolter, and V. Koltun, “An empirical evaluation of generic convolutional and recurrent networks for sequence modeling,” *arXiv preprint arXiv:1803.01271*, 2018.
- [10] S. Becker, R. Hug, W. Hübner, and M. Arens, “An evaluation of trajectory prediction approaches and notes on the trajnet benchmark,” *arXiv preprint arXiv:1805.07663*, 2018.
- [11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [12] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, “Social-stgcn: A social spatio-temporal graph convolutional neural network for human trajectory prediction,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 424–14 432.
- [13] A. Vemula, K. Muelling, and J. Oh, “Social attention: Modeling attention in human crowds,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018, pp. 4601–4607.

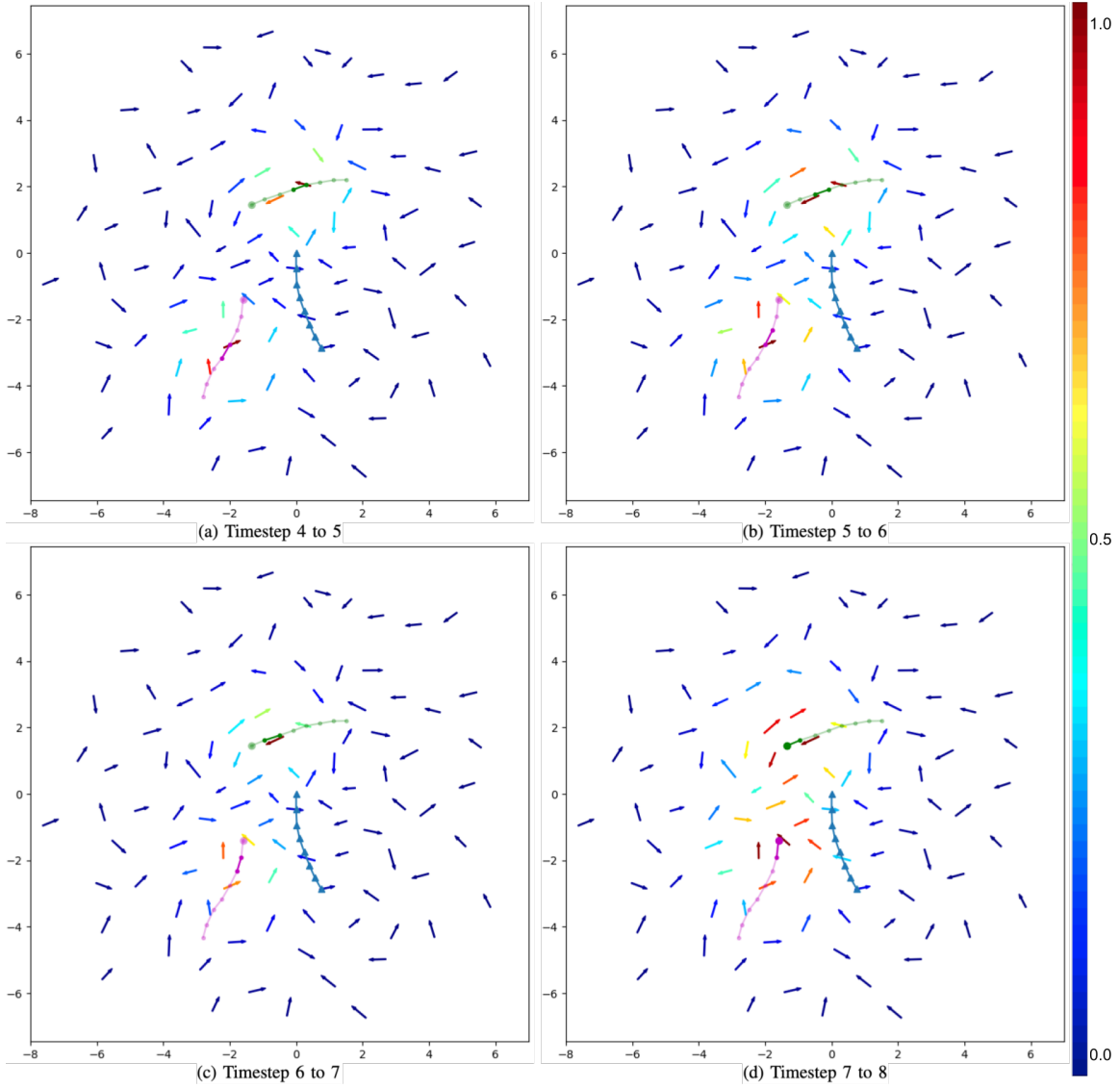


Fig. 7. “Noticing motion patterns” when presented trajectory data. The arrows are from Fig.6(c). All trajectories are observed history. Dots are context trajectories and larger round dots indicate the last-timestep locations. Blue triangles are the target trajectory which is only plotted but not matched to the learned patterns because patterns here are for context trajectories only. Trajectories are 8-timestep long, but the first 3 timesteps are neglected. Each subgraph demonstrates the matching of the particular trajectory segment, e.g., (d) shows the last segment, (c) shows the 2nd last, etc. By Pattern Extraction Convolution (PEC) layer, the segment is matched to similar patterns, causing these “noticed patterns” to be shown in heated color.

- [14] V. Kosaraju, A. Sadeghian, R. Martín-Martín, I. Reid, H. Rezatofighi, and S. Savarese, “Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks,” in *Advances in Neural Information Processing Systems*, 2019, pp. 137–146.
- [15] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras, “Human motion trajectory prediction: A survey,” *The International Journal of Robotics Research*, vol. 39, no. 8, pp. 895–935, 2020.
- [16] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in neural information processing systems*, 2019, pp. 8026–8037.
- [17] L. Zhang, Q. She, and P. Guo, “Stochastic trajectory prediction with social graph network,” *arXiv preprint arXiv:1907.10233*, 2019.
- [18] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, “You’ll never walk alone: Modeling social behavior for multi-target tracking,” in *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 2009, pp. 261–268.
- [19] 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.
- [20] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [21] X. Yao, J. Zhang, and J. Oh, “Following social groups: Socially compliant autonomous navigation in dense crowds,” *arXiv preprint arXiv:1911.12063*, 2019.