RNN-based Prediction of Pedestrian Turning Maneuvers

Stefan Becker

Fraunhofer Institute for Optronics,
System Technologies, and Image Exploitation IOSB
Gutleuthausstr. 1, 76275 Ettlingen, Germany
stefan.becker@iosb.fraunhofer.de

Technical Report IES-2018-05

Abstract

The dynamics of objects, such as pedestrians, varies over time. Commonly this problem is tackled with traditional approaches like the Interacting Multiple Model (IMM) filter using a Bayesian formulation. Following the current trend towards using deep neural networks, in this paper an RNN-based alternative solution for pedestrian maneuver prediction is presented. Similar to an IMM filter solution, the presented model assigns a confidence value to a performed dynamic and, based on them, puts out a multi-modal distribution over future pedestrian trajectories. The qualitative evaluation is done on synthetic data, reflecting prototypical pedestrian maneuvers.

1 Introduction

The applications of pedestrian path prediction cover a wide range from robot navigation, autonomous driving, smart video surveillance to object tracking. Traditionally, the task of object motion prediction is addressed by using a Bayesian Formulation in approaches such as the Kalman filter [Kal60], or nonparametric methods, such as particle filters [AMGC02]. Following the success of recurrent neural networks (RNNs) in modeling temporal dependencies in a variety

of sequence processing tasks, such as speech recognition [GMH13, CKD+15] and caption generation [DHG+15, XBK+15], RNNs are increasingly utilized for object motion prediction [AGR+16, ARG+17, HBHA17, HBHA18, BHHA18]. When relying on traditional approaches, the challenge of changing dynamics over time or rather maneuvers is commonly done with the Interacting Multiple Model (IMM) filter [BBS88]. The IMM filter is an elegant way to combine a set of candidate models into a single context by weighting each individual model. Each model corresponds to a specific motion pattern and contributes to the final state estimation depending on its current weight. According to the IMM filter solution, in this paper an RNN-based model is presented, which on the one hand is able to also provide a confidence value for the performed dynamic and on the other hand can overcome some limitations of the IMM filter. The suggested RNN-encoder-decoder model generates the probability distribution over future pedestrian paths conditioned on a maneuver class. The model is based on the work of Deo and Trivedi [DT18]. For the case of freeway traffic, they used an RNN-encoder-decoder network for vehicle maneuver and trajectory prediction. In the context of vehicle motion prediction, maneuver classes can be better defined than for pedestrians. Due to the dynamic behavior of pedestrians, the maneuver classes are here defined based on the deviation of a straight walking pedestrian. The presented network adapts the maneuver network of Deo and Trivedi with insights of the work of Becker et al. [BHHA18] for RNN-based pedestrian trajectory prediction. The analysis is done on synthetic data reflecting prototypical scenarios capturing turning maneuvers of pedestrians.
In the following, a brief formalization of the problem and a description of the RNN-based model are provided. The qualitative achieved results are presented in section 3. Finally, a conclusion is given in section 4.

2 RNN-based Pedestrian Maneuver Prediction

The goal is to devise a model that can successfully predict future paths of pedestrians and represent alternating pedestrian dynamics, e.g. dynamics that can transition from a straight walking to a turning maneuver. Here, trajectory prediction is formally stated as the problem of predicting the future trajectories of a pedestrian, conditioned on its track history. Given an input sequence $\mathcal{O} = \{(x^t, y^t) \in \mathbb{R}^2 | t = 1, \ldots, t_{obs}\}$ of T_{obs} consecutive pedestrian positions $\vec{x}^t = (x^t, y^t)$ at time

t along a trajectory the task is to generate a multi-modal prediction for the next T_{pred} positions $\{\vec{x}^{t+1}, \vec{x}^{t+2}, \ldots, \vec{x}^{t+T_{pred}}\}$. One insight of the work Becker et al. [BHHA18] is that motion continuity is easier to express in offsets or velocities, because it takes considerably more modeling effort to represent all possible conditioning positions. For exploiting scene-specific knowledge for trajectory prediction, additional use of the position information is required. When sufficient training samples from a particular scene are available, Hug et al.[HBHA17] showed that RNN-based trajectory prediction models are able to capture spatially depending behavior changes only from motion data. However, here the offsets are used for conditioning the network $\mathcal{O} = \{(\delta_x^t, \delta_y^t) \in \mathbb{R}^2 | t = 2, \ldots, t_{obs}\}$. Apart from the smaller modeling effort to represent conditioned offsets, the shift to offsets helps to prevent undefined states due to a limited data range [BHHA18]. Furthermore, it is easier to capture the scene-independent aspect of human behavior and to better generalize across datasets. The future trajectory is denoted with $\mathcal{Y} = \{(x^t, y^t) \in \mathbb{R}^2 | t = t_{obs} + 1, \ldots, t_{pred}\}$ and the model estimates the conditional distribution $P(\mathcal{Y}|\mathcal{O})$. In order to identify specific dynamics under M desired maneuver classes (e.g. turning maneuvers and straight walking), this term can be given by:

$$P(\mathcal{Y}|\mathcal{O}) = \sum_{i}^{M} P_{\Theta}(\mathcal{Y}|m_i, \mathcal{O}) P(m_i|\mathcal{O})$$

Here, $\Theta = \{\Theta^{t_{obs}+1}, \ldots, \Theta^{t_{pred}}\}$ are the parameters of a L component Gaussian mixture model $\Theta^t = (\vec{\mu}_l^t, \Sigma_l^t, w_l^t)_{l=1,\ldots,L}$. By adding the maneuver context in form of the posterior mode probability, $P(m_i|\mathcal{O}) \stackrel{\wedge}{=} \alpha_i$ the analogy to the classic IMM filter becomes apparent. For an IMM filter the mode probability is used to calculate the mixing probabilities to combine the set of chosen candidate models into a merged estimate. In case of using an IMM filter the time behavior of the basic filter set is modeled as a homogeneous (time invariant) Markov chain with a fixed transition probability matrix (TPM) $m_{ij} \stackrel{\wedge}{=} P(m_i^t | m_j^{t-1})$. Instead of setting the parameter of the time behavior manually, the current mode probability is inferred from the hidden states of the RNN. For the proposed RNN-based pedestrian maneuver prediction model, the basic architecture is a Recurrent-Encoder-Decoder model. The encoder takes the frame by frame input sequence $\mathcal{O}$. The hidden state vector of the encoder is updated at each time step based on the previous hidden state and the current offset. The generated internal representation

is used to predict mode probability $\vec{\alpha}^t$ at the current time step. The encoder can be defined as follows:

$$\begin{aligned}\vec{h}^t_{encoder} &= \text{RNN}(\vec{h}^{t-1}_{encoder}, \vec{\delta}^t_{(x,y)}; W_{encoder}) \\ \vec{\alpha}^t_{logits} &= \text{MLP}(\vec{h}^t_{encoder}; W_{en}) \\ \hat{\vec{\alpha}}^t &= \frac{\exp\left(\vec{\alpha}^t_{logits}\right)}{\sum_{j=1}^{M} \exp\left(\alpha^t_{logits,j}\right)}\end{aligned}$$

Here, $RNN(\cdot)$ is the recurrent network, $\vec{h}$ the hidden state of the RNN and $MLP(\cdot)$ the multilayer perceptron. W represents the weights and biases of the MLP or respectively RNN. The final state of the encoder can be expected to encode information about the track histories. For generating a trajectory distribution over dynamic modes, the encoder hidden state is appended with a one-hot encoded vector corresponding to specific maneuvers. Hence the network is conditioned purely on offsets, position information is required to localize. Localization information persists here only implicitly by performing path integration and using the last observed point $\vec{x}^{t_{obs}}$ as reference point. The decoder of the model can be defined as follows:

$$\begin{aligned}\vec{h}^t_{decoder} &= \text{RNN}(\vec{h}^{t-1}_{decoder}[\vec{h}^t_{encoder}], \vec{\alpha}^t; W_{decoder}) \\ \hat{\mathcal{Y}} = \{(\hat{\vec{\mu}}^t_l + \vec{x}^{t_{obs}}, \hat{\Sigma}^t_l, \hat{w}^t_l) | t = t_{obs}+1, \ldots, t_{pred}\} &= \text{MLP}(\vec{h}^t_{decoder}; W_{de})\end{aligned}$$

The decoder is used to parametrize a mixture density output layer (MDL) or rather Θ directly for several positions in the future. Nevertheless, the overall RNN-based pedestrian maneuver prediction network uses the trajectory prediction and dynamic classification jointly, the loss function for training is splitted into two parts. Dynamic classification is trained to mimimize the sum of cross-entropy losses of the different M motion model classes:

$$\mathcal{L}(\mathcal{O})_{maneuver} = -\sum_{j=1}^{M} \alpha^t_{j,GT} \log(\hat{\alpha}^t_j)$$

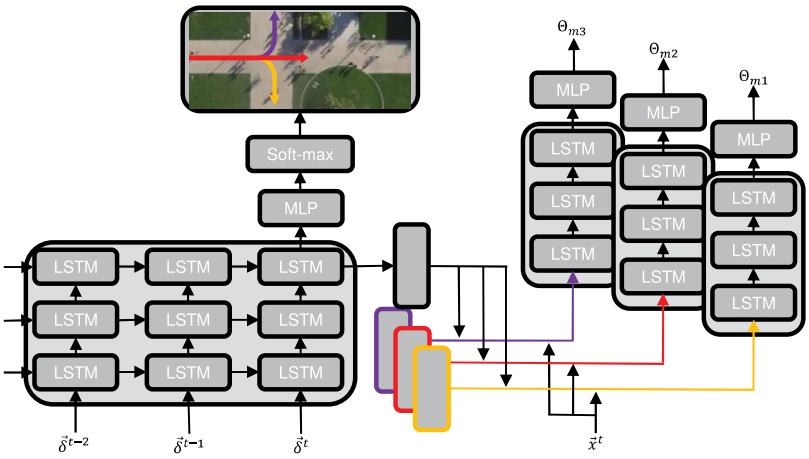

Figure 2.1: Visualization of the RNN-encoder-decoder network for jointly predicting specific dynamic probabilities and corresponding future distributions of trajectory positions. The encoder predicts the dynamic or rather maneuver probabilities and the decoder uses the context vector to predict future pedestrian locations.

Additionally, the encoder-decoder is trained by minimizing the negative log likelihood for the ground truth future pedestrian locations conditioned under the performed maneuver class. The context vector is appended with the ground truth values of the maneuver classes for each training trajectory. This results in the following loss function:

$$\mathcal{L}(\mathcal{O})_{pred} = -\log(P_\Theta(\hat{\mathcal{Y}}|m_{GT}, \mathcal{O})P(m_{GT}|\mathcal{O}))$$
$$\mathcal{L}(\mathcal{O})_{pred} = \sum_{t=t_{obs}+1}^{t_{pred}} -\log(\sum_{l=1}^{L} \hat{w}_l^t \mathcal{N}(\vec{x}^t|\hat{\vec{\mu}}_l^t + \vec{x}^{t_{obs}}, \hat{\Sigma}_l^t; m_{GT}))$$

The overall architecture is visualized in figure 2.1. The context vector combines the encoding of the track history with the encoding of the alternating dynamic classes and is used as input for the decoder.

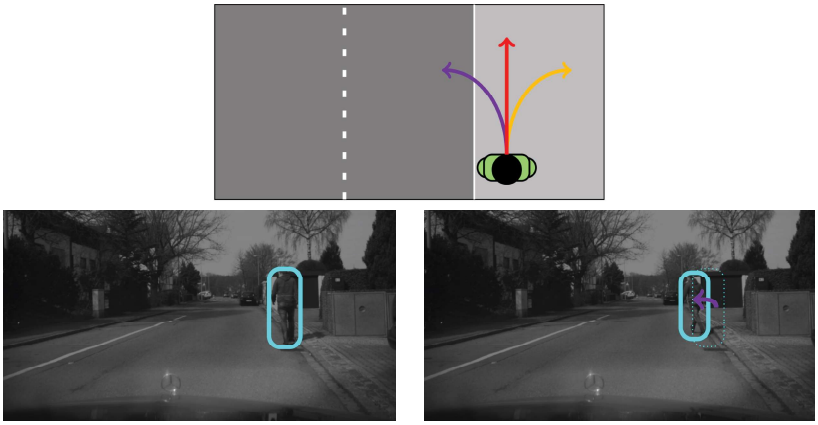

Figure 3.1: Illustration of a typical pedestrian motion. The above image images depicts the three chosen maneuver classes of straight walking and taking a turn to the left or the right. The images below show a person changing from straight walking to crossing the street complying to a change from the defined maneuver classes. In particular from straight walking to turning left. In the context of intelligent vehicles, this is often called bending in [SG13].

3 Data Generation and Evaluation

This section consists of a brief qualitative evaluation of the proposed approach. The evaluation is concerned with verifying the overall viability of the approach in maneuver situations. For initial results, a synthetic test condition is used in order to gain insight into the model behavior in different typical pedestrian motion types. A prototypical maneuver performed by a pedestrian with keen interest in the context of intelligent vehicles and video surveillance is a turning maneuver. In such a maneuver the dynamics of a pedestrian changes from a straight walking into a bending in behavior. So long as a person moves in a straight line at a reasonably constant speed, its dynamics can be captured with a Kalman filter and a constant velocity model. During the maneuver, the relation to one fixed process model describing the dynamics fails. In the context of modeling the dynamics of pedestrians this switch in dynamics is normally modeled with an IMM filter. For example in the work of Schneider et al. [SG13] or Kooij et al. [KSFG14], the motion of pedestrians is modeled with an IMM filter combining basic models like constant velocity and constant acceleration model. In such a situation the

motion changes from a rectilinear dynamic to a curvilinear motion, in relation to the dynamic this results in an additional acceleration. Therefore, a change from a constant velocity model to a turning model or acceleration model indicates a critical situation from the vehicle perspective. Figure 3.1 illustrates such a turning maneuver.

For generating synthetic trajectories of a basic maneuvering pedestrian, random agents are sampled from a Gaussian distribution according to the preferred pedestrian walking speed [Tek02] ($\mathcal{N}(1, 38\frac{m}{s}, 0.37\frac{m}{s})$). During a single trajectory simulation the agents can perform a turning maneuver. For the presented results the turning event takes 5 steps for a $90°$ change in heading with a fixed frame rate of 1 frame per second. The observation noise of the position sensor is assumed to be Gaussian distributed in x and y with $\sigma = 0.2m$. As mentioned above, a definition of maneuver class for pedestrians is harder to establish than for vehicles. Here, the main interest is here to detect a deviation from a standard behavior, and whether the pedestrian is in a *normal* mode. A set deviation in heading for a required time horizon can then be used to assign maneuver labels to single trajectories. As the distribution over the trajectories is captured with a Gaussian mixture model the maneuver description for the outlier trajectory distribution can still be multi-modal. For the *normal* or straight motion a single Gaussian component is sufficient. In case of the generated synthetic data, the turning maneuver trajectory distribution could be captured using one Gaussian component.

The model has been implemented using *Tensorflow* [Aba15] and is trained for 300 epochs using ADAM optimizer [KB15] with a fixed learning rate of 0.003. For the experiments the RNN variant Long Short-Term Memory [HS97] (LSTM) is used. In figure 3.2 predictions for three different preformed motion types are depicted. In all shown images the maneuver has started two time steps before. The resulting multi-modal prediction is visualized as a heatmap for the images on the left. On the right, the visualization shows the predicted covariances for 12 future positions weighted by the predicted maneuver probability and temporally. Turning to the left is highlighted in purple, walking straight with red and turning to the right in yellow.

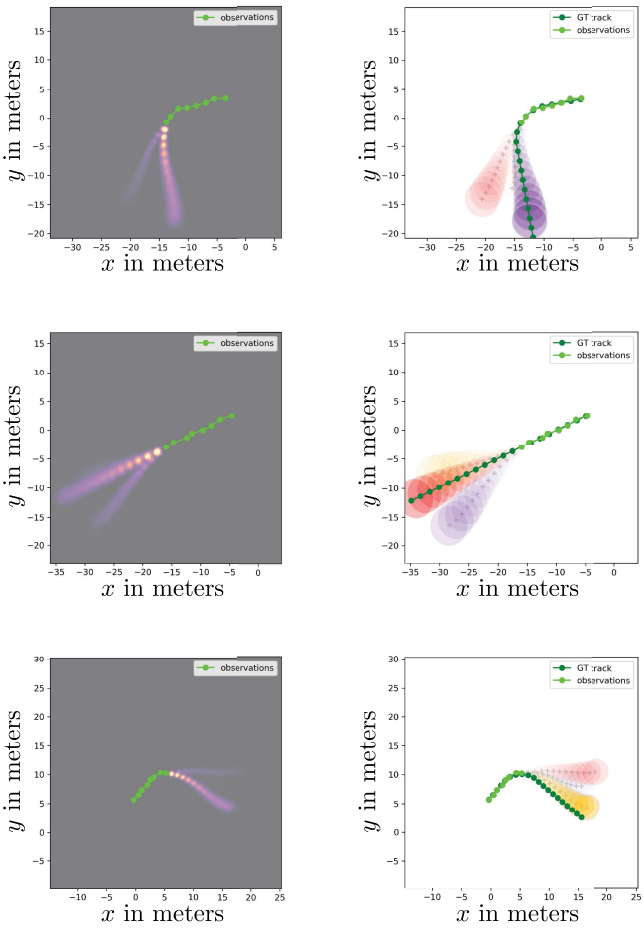

Figure 3.2: Visualization of the multi-modal predictions of the network. (Left) Density plots of the three different maneuvers. (Right) Visualization of the predicted covariance matrices with a temporal weighting and depending on the estimated maneuver probability. The turning maneuver predictions are visualized in purple and respectively yellow. The predictions for straight walking are highlighted in red.

The shown results are achieved based on noisy observations. The model is able to successfully recognize the turning behavior and to produce a reasonable distribution for the further positions. Without the explicit splitting into maneuver classes, an RNN-based solution which generates a Gaussian mixture model condition on the input sequence, is also capable to produce a similar multi-modal distribution (see for example the work of Hug et al. [HBHA17]). However, the presented model is able to successfully assign probabilities to current performed pedestrian behavior instead of only encoding this information in the hidden states of the RNN. Similar to the provided mode probabilities of IMM filters this can be used for further processing steps. Thus, the presented RNN-based model is able to also provide a confidence value $P(m_i|\mathcal{O}) \triangleq \alpha_i$ for the performed dynamic, but to avoid modeling the dynamic transitions with a fixed transition probability matrix $P(m_i^t|m_j^{t-1})$. Further, instead of choosing the basic filter set, the prediction model is learned. In case there exists some well known model for describing the standard dynamic of the desired target, only deviations from the known dynamic can be used to define additional maneuver classes.

4 Conclusion

In this report, an RNN-encoder-decoder model aimed to jointly predicting specific dynamic probabilities and corresponding distributions of future pedestrian trajectory has been presented. The model capabilities were shown on synthetic data reflecting typical pedestrian maneuvers. By conditioning on specific dynamic models or rather deviation of standard behavior, the model makes it possible to generate additional information in terms of an assigned maneuver probability similar to an IMM filter, but without the explicit modeling of the dynamic transitions.

Bibliography

[Aba15] M. Abadi et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[AGR+16] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social LSTM: Human trajectory prediction in crowded spaces. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–971. IEEE, 2016.

[AMGC02] M.S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *Transactions on Signal Processing*, 50(2):174–188, 2002.

[ARG+17] A. Alahi, V. Ramanathan, K. Goel, A. Robicquet, A. Sadeghian, L. Fei-Fei, and S. Savarese. Learning to predict human behaviour in crowded scenes. In *Group and Crowd Behavior for Computer Vision*. Elsevier, 2017.

[BBS88] H.A.P. Blom and Y. Bar-Shalom. The interacting multiple model algorithm for systems with markovian switching coefficients. *Transactions on Automatic Control*, 33(8):780–783, 1988.

[BHHA18] S. Becker, R. Hug, W. Hübner, and M. Arens. Red: A simple but effective baseline predictor for the trajnet benchmark. In *European Conference on Computer Vision (ECCV) Workshops*. Springer International Publishing, 2018.

[CKD+15] J. Chung, K. Kastner, L. Dinh, K. Goel, A. Courville, and Y. Bengio. A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[DHG+15] J. Donahue, L.A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015.

[DT18] N. Deo and M.M. Trivedi. Multi-modal trajectory prediction of surrounding vehicles with maneuver based lstms. In *Intelligent Vehicles Symposium (IV)*, pages 1179–1184. IEEE, 2018.

[GMH13] A. Graves, A.R. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649, 2013.

[HBHA17] R. Hug, S. Becker, W. Hübner, and M. Arens. On the reliability of LSTM-MDL models for pedestrian trajectory prediction. In *Representations, Analysis and Recognition of Shape and Motion from Imaging Data (RFMI)*, Savoie, France, 2017.

[HBHA18] R. Hug, S. Becker, W. Hübner, and M. Arens. Particle-based pedestrian path prediction using LSTM-MDL models. In *21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2684–2691, 2018.

[HS97] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997.

[Kal60] R.E. Kalman. A new approach to linear filtering and prediction problems. *ASME Journal of Basic Engineering*, 82, 1960.

[KB15] D.P. Kingma and J. Ba. Adam: A method for stochastic optimization. *International Conference for Learning Representations (ICLR)*, 2015.

[KSFG14] J.F.P. Kooij, N. Schneider, F. Flohr, and D.M. Gavrila. Context-based pedestrian path prediction. In *European Conference on Computer Vision (ECCV)*, pages 618–633. Springer International Publishing, 2014.

[SG13] N. Schneider and D.M. Gavrila. Pedestrian path prediction with recursive bayesian filters: A comparative study. In *German Conference on Pattern Recognition (GCPR)*, pages 174–183. Springer Berlin Heidelberg, 2013.

[Tek02] K. Teknom. *Microscopic Pedestrian Flow Characteristics: Development of an Image Processing Data Collection and Simulation Model*. PhD thesis, Tohoku University, 2002.

[XBK+15] 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 *International Conference on Machine Learning (ICML)*, volume 37, pages 2048–2057, Lille, France, 2015. PMLR.