Learning Pedestrian Crowd as Clusters using Abstracted Features

Yasunori Yokojima[1] , **Tatsuhide Sakai**[2]
[1]Siemens K.K.
[2]DAIHATSU MOTOR CO. LTD.
yasunori.yokojima@siemens.com, Tatsuhide_Sakai@dk.daihatsu.co.jp

Abstract

In this study, we propose to apply deep learning to pedestrian crowd represented as clusters, learn to predict their dynamics by a neural network, and examine the learned contents in terms of the cognitive-level analogy. By regarding the trained model as a model at a certain cognitive level, the model can be applied to distinguish phenomenons that learned in the current cognitive level to the others requiring further details of surroundings. To demonstrate this methodology, we introduce a neural network and training data recorded at public locations and examine the trained model. We aim to contribute to developing safe and secure traffic systems as an application of the proposed approach.

1 Introduction

As in deep learning and optimization techniques, it is often useful to look at biological analogies in developing new technologies. The concepts built on such biological analogies can be a basis for the technologies that follow. In this study, we propose a new approach to model pedestrian crowd dynamics based on our observations of cognitive modes.

The modeling of pedestrian behavior has long been a subject of study. In particular, the recent advancements by data-driven approaches [Alahi *et al.*, 2016] [Gupta *et al.*, 2018][Vemula *et al.*, 2018] are more suitable for the objectives of predicting independent pedestrian trajectories than conventional dynamic modeling approaches [Helbing and Molnar, 1995]. Typically in data-driven approaches, features of pedestrians are fed to recurrent units, and interactions are taken into account by interconnections, such as pooling layers. In this configuration, features for pedestrians need to be measured accurately.

In our daily cognition, there are two modes: the first mode corresponds to a clear perception of surroundings and conscious decision, and the second mode corresponds to a less conscious perception of pedestrians as clusters and quick desitions. Humans efficiently behave while switching between these cognitive modes depending on the situation. We infer the existence of such cognitive modes suggests that there should be a set of cognitive models selectable depending on the objectives of applications.

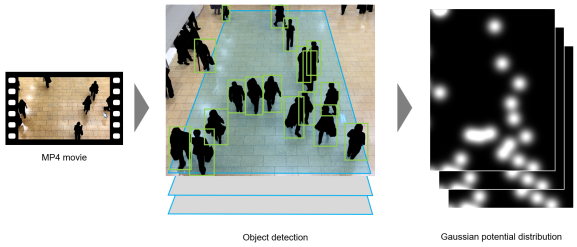

Figure 1: Training Data Preparation

Based on this observation, we propose an approach to regard pedestrians as clusters as a complement to the existing data-driven approaches, which require detailed features of independent pedestrians. By regarding a cluster of people as a subject of study, it enables us to model a system based on features of clusters instead of features of agents, whose maximum numbers are inherently limited by the neural network size. In the cluster-based approach, we expect interactions within a cluster are learned through training once we set up the data and learning appropriately.

2 Training data and Augumentation

For training a network, minutes of movies were recorded by a monocular camera in public locations, including a pedestrian exclusive road and exhibition booth. We extracted frames from these movies, and applied the object detection to detect pedestrians on frames [Redmon and Farhadi, 2018]. Approximate positions of pedestrians are defined using the bounding boxes. We converted them into points on a 2-d rectangle via a frame conversion using four corner markers on the ground (Figure 1).

Based on computed pedestrian positions on a rectangle, we generated input images for training and validation using a Gaussian potential function [Saiin *et al.*, 2018]. We set the center of the Gaussian potential to a pedestrian position and the standard deviation to 0.5 [m], which is in accordance with a typical personal distance defined in proxemics.

By adding up Gaussian functions, it results in a 2-d Gaussian mixture distribution, which can be rendered as blobs on black canvas, as shown in Figure 1.

These images are randomly cropped as 128×128-pixel im-

ages. The actual area size corresponding to this pixel size depends on the original image size.

3 The neural network architecture

To demonstrate learning the temporal changes of clusters, we consider a network shown in Figure 2. The network has encoder converting input images to high-level feature representations. For sequential images, we applied the same encoder multiple times to prepare a time series of high-level features, which is similar to the Siamese network [Taigman *et al.*, 2014]. Then, these features are fed as inputs to the LSTM layer to predict the next features. The predicted features are converted into an image by the decoder network.

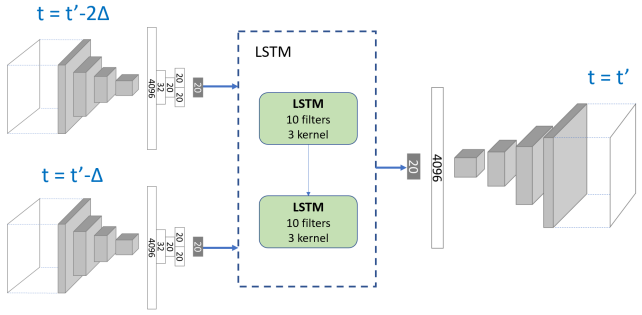

Figure 2: Encoder-LSTM-Decoder architecture

By reflecting the probabilistic aspect of pedestrians, the encoder and decoder are pre-trained as β-VAE to learn a static pedestrian distribution in the training data [Higgins *et al.*, 2017] [Burgess *et al.*, 2018]. After the pre-training, we constructed the encoder-LSTM-decoder architecture by transferring the encoder and decoder with pre-trained weights, as in Figure 2.

4 Prediction

By feeding two sequential frames to the trained network, it can predict the next frame. As networks trained by assuming a specific sampling rate, prediction always keeps the same sampling rate unless the network is re-trained with the other value.

Similar to a text sequence generation using a recurrent neural network, a sequence can be generated by iteratively feeding the predicted image, as shown in Figure 3. The network first receives two sequential frames as seeds and predicts the next frame, and this prediction becomes one of the inputs to the next prediction. We repeated these steps over a sequence, as shown from left to right in Figure 3.

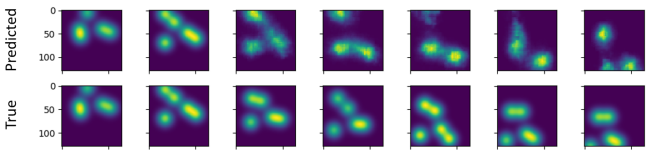

Figure 3: Sequence Generation

5 Discussions

5.1 Cognitive level and Prediction

As demonstrated in Section 4, the moves of pedestrian clusters are learned and successfully predicted using abstracted features. Since the network learned mean phenomena in samples at every 3 FPS, it is not capable of predicting specific actions noticeable to human eyes, such as a person walking against a mean traffic flow and quickly moving to avoid contact. We assume such a quick move represents awareness to detailed features of other pedestrians, i.e., high cognitive mode, which is now distinguished as an error to the prediction by our model.

For lifting the cognitive level, it would be useful to introduce more features and annotate them in the training data, such as pedestrians face directions, gaze, pose, etc., and attention mechanism [Vemula *et al.*, 2018] to set focus on features. Towards this direction, the latest results are available in the existing data-driven approaches [Alahi *et al.*, 2016] [Gupta *et al.*, 2018].

For the temporal aspect, since we used only two frames to predict the next frame, the model can only make a short-term prediction. For predicting a long-term correlation, we can increase the number of input frames, and expect LSTM layers to capture the long-term correlation of high-level features. We also infer pedestrian actions inherently have multiple time-scale nature and switching mechanisms, depending on a circumstance.

5.2 Pedestrian dynamics and Environment

We applied the current framework to movies recorded at two different locations. The first set of movies recorded at a public concourse has a trend of vertical pedestrian flow. This trend is learned well and reflected in the prediction. On the other hand, the movies recorded at an exhibition booth show a pattern of pedestrians to stop and stay at some locations. This trend contributed negatively to training, and prediction performance was slightly lower than the former.

5.3 Learning using Abstracted Features

Learning on abstract features using the encoder-LSTM-decoder structure brings some advantages against pixel-to-pixel image prediction [Shi *et al.*, 2015] [Mahjourian *et al.*, 2017] [Sakurai *et al.*, 2019]. Firstly, the model training becomes less computationally intensive because of the smaller dimensionality. Secondly, the network trained on high-level features is robust to small noises in inputs [Villegas *et al.*, 2017]. Thirdly, pre-trained weights can facilitate the training of derived task, which is the case in our two-step training explained in Section 3

6 Conclusion

We proposed a deep learning application to study pedestrian crowds as clusters based on an analogy of cognitive modes. We hope to contribute to design safe and secure traffic systems as an application of the cognitive modeling.

Acknowledgments

We are grateful to Mr. Balakrishnan Ayyanar (Siemens K.K.) for valuable comments and suggestions.

References

[Alahi *et al.*, 2016] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. *IEEE Conference on Conputer Vision and Pattern Recognition (CVPR)*, pages 961–971, 2016.

[Burgess *et al.*, 2018] Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexaner Lerchner. Understanding disentangling in β-vae. *CoRR, abs/1804.03599*, 2018.

[Gupta *et al.*, 2018] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12126–12134, 2018.

[Helbing and Molnar, 1995] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical Review E*, 51(5):4282, 1995.

[Higgins *et al.*, 2017] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-vae: Learning basic visual concepts with a constrained variational framework. *ICLR*, 2017.

[Mahjourian *et al.*, 2017] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Geometry-based next frame prediction from monocular video. *IEEE Intelligent Vehicles Symposium (IV)*, pages 1700–1707, 2017.

[Redmon and Farhadi, 2018] Joseph Redmon and Ali Farhadi. Yolov3: an in-cremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[Saiin *et al.*, 2018] Ryuji Saiin, Tomohiro Daimon, Takayoshi Yamashita, and Masahiko Nakamura. Multiple object motion prediction using deep convolutional neural networks. *Transactions of the Society of Automotive Engineering of Japan*, 49(2), 2018.

[Sakurai *et al.*, 2019] Shunsuke Sakurai, Hideaki Uchiyama, Atshushi Shimada, and Rin ichiro Taniguchi. Plant growth prediction using convolutional lstm. *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2019.

[Shi *et al.*, 2015] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai kin Wong, and Wang chun Woo. Convolutional lstm network: a machine learning application for precipitation nowcasting. *Advances in Neural Information Processing Systems*, pages 802–810, 2015.

[Taigman *et al.*, 2014] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: closing the gap to human-level performance in face verification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014.

[Vemula *et al.*, 2018] Anirudh Vemula, Katharina Muelling, and Jean Oh. Social attention: Modeling attention in human crowds. *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–7, 2018.

[Villegas *et al.*, 2017] Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. Learning to generate long-term future via hierarchical prediction. *International Conference on Machine Learning (ICML)*, 2017.