Using occupancy priors to generalize people flow predictions

Francesco Verdoja, Tomasz Piotr Kucner, Ville Kyrki

***Abstract*— Mapping people dynamics is a crucial skill for robots, because it enables them to coexist in human-inhabited environments. However, learning a model of people dynamics is a time consuming process which requires observation of large amount of people moving in an environment. Moreover, approaches for mapping dynamics are unable to transfer the learned models across environments: each model is only able to describe the dynamics of the environment it has been built in. However, the impact of architectural geometry on people's movement can be used to anticipate their patterns of dynamics, and recent work has looked into learning maps of dynamics from occupancy. So far however, approaches based on trajectories and those based on geometry have not been combined. In this work we propose a novel Bayesian approach to learn people dynamics able to combine knowledge about the environment geometry with observations from human trajectories. An occupancy-based deep prior is used to build an initial transition model without requiring any observations of pedestrian; the model is then updated when observations become available using Bayesian inference. We demonstrate the ability of our model to increase data efficiency and to generalize across real large-scale environments, which is unprecedented for maps of dynamics.**

I. Introduction

Over the past years, we have observed the development of different methodologies for modelling pedestrian motion, one of which is maps of dynamics (MoDs) [1]. MoDs capture the common motion patterns followed by uncontrolled agents (*i.e.*, humans, human-driven vehicles) in the environment and enable robots to anticipate typical behaviors throughout the environment. Unfortunately, the process of building MoDs is very time- and resource-consuming: reliable MoDs are built through measuring repeating motion patterns executed by uncontrolled agents in the given environment. As a consequence, the deployment of a successful robotic system using MoDs requires a substantial amount of time necessary to collect enough relevant data [1], [2]. Moreover, a MoD is only able to describe and predict pedestrian motion in the same environment it has been built in. The inability to transfer between environments is a crucial limitation of MoDs, especially considering that pedestrian traffic rules share commonalities across environments, *e.g.*, people move similarly through a corridor, or around a door.

To address this limitation, recently we have seen the development of methods that leverage the correlation between the shape of the environment and the behavior of humans therein, to predict the possible motion patterns in it. That said, the existing efforts have been narrow in scope, primarily focusing either on the use of synthetic trajectory data [3], [4], or being limited to small environments [5], [6]. Moreover, how to combine these occupancy-based methods with observations of human movement has not yet been studied.

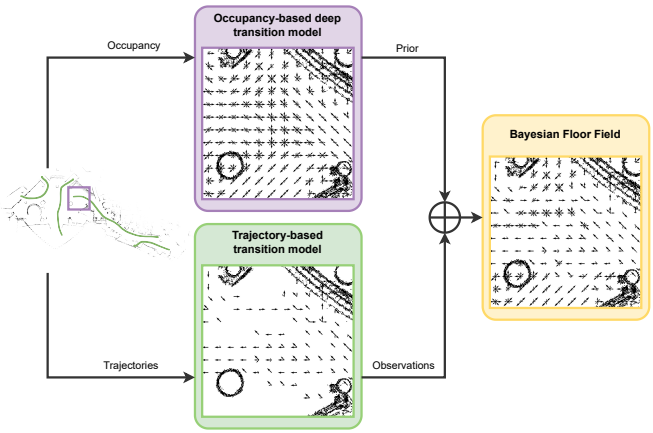


Fig. 1: We propose Bayesian Floor Field, a method to build maps of human dynamics by combining information from static occupancy and pedestrian trajectories.

In this work we propose a novel Bayesian MoD approach using people dynamics learned from environment occupancy as prior, and updating the model with measured human trajectories. An illustration of the approach is given in Fig. 1.

In particular, the contributions of this work are:

1) A novel approach for training occupancy-based deep transition probability models from real data;
2) A novel method for building MoDs by combining knowledge about the environment geometry with observations of people motion;
3) A study over the ability of the proposed occupancy-based approach to model real unseen pedestrian motion both in the same environment it was trained on, as well as in a completely different large-scale environment never seen during training;
4) Experimental evidence that knowledge about the environment occupancy can reduce the amount of trajectory data required to build MoDs, by comparing the proposed method against a traditional MoD approach.

II. Related work

The observation that people tend to follow spatial or spatio-temporal patterns enabled the development of MoDs.

This work was supported by the Research Council of Finland, decision 354909. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

F. Verdoja, T. Kucner and V. Kyrki are with School of Electrical Engineering, Aalto University, Finland. (name.surname@aalto.fi)

MoDs are a special case of semantic maps, where information about motion patterns is retained as a feature of the environment [1]. In this work, we are especially focusing on directional MoDs, which are well suited to represent local dynamic patterns caused by directly observed moving agents while being robust against partial or noisy observations [2], [7]–[9]. Furthermore, this type of MoDs consists of a large spectrum of representations of varying levels of expressiveness and complexity. Including fairly simple models such as floor fields [10] as well as more complex multimodal, continuous representations [11].

Even though the idea of utilizing metric information to inform dynamics has substantially impacted the motion prediction community it has not yet received adequate attention in the field of MoDs. One of the more impactful attempts in this direction is the work by Zhi *et al.* [4]. In that work, the authors utilize artificially generated trajectories to train a deep neural network to predict possible behaviors in new unobserved environments. At this same time, Doelinger *et al.* [5], [6] proposed a method to predict not the motion itself but the levels of possible activity in given environments, based on surrounding occupancy. Both works by Zhi *et al.* [4] and Doelinger *et al.* [5], [6] have limitations by either being application specific and narrow in scope, using either only synthetic data, or by being limited to only small environments. In this work, we propose a step change with respect to the presented state-of-the-art by presenting a way to leverage real human trajectories in large-scale environments and open new possibilities for predicting not only motion patterns but other environment-dependent semantics. Moreover, we present the first approach to combine occupancy-based and trajectory-based MoDs.

III. Method

Let us represent the environment $\mathbf{M}$ the robot is operating in as a collection of cells $c \in \mathbf{M}$. For each cell c, we assume to know an occupancy probability $s(c) \in [0, 1]$ describing the likelihood of that portion of the environment to be occupied by static objects. We refer to s as the *static occupancy map* of the environment built following [12].

To model people movement in the environment, for each cell c, we want to determine the likelihood that a person in c will head in a particular direction $\delta \in [0, 2\pi)\,\mathrm{rad}$. Formally, we define the transition model for cell c as a categorical distribution $\mathrm{Cat}(k, \mathbf{d}_c)$ over k discrete directions equally dividing the range $[0, 2\pi)\,\mathrm{rad}$, *i.e.*,

$$\mathsf{P}(\delta \mid \mathbf{d}_c) = \sum_{i=1}^{k} d_{ic} \mathbf{1}_i(\delta) \ , \tag{1}$$

where $\mathbf{d}_c = (d_{1c}, \ldots, d_{kc} \mid \sum_{i=1}^{k} d_{ic} = 1)$, d_{ic} represents the probability of moving toward direction i from c, and the indicator function $\mathbf{1}_i(\delta) = 1$ iff $\frac{2(i-1)}{k}\pi \leq \delta < \frac{2i}{k}\pi$, 0 otherwise. We use $\mathbf{d}_c$ as shorthand for $\mathbf{d}(c)$. We refer to the complete model $\mathbf{d}$ as the *map of dynamics* (or *people flow map*) of the environment.

Previous works building MoDs as grid-based categorical distributions infer the distribution parameter $\mathbf{d}_c$ for each cell c from trajectory data in a frequentistic fashion [2], [10]. Instead, in this work we treat it as a Bayesian inference problem.

The main assumptions in this work are that: (i) an environment's occupancy around a certain location (*i.e.*, that location's neighborhood) influences how people move from it; and (ii) neighborhoods having similar occupancy, even from different environments, influence people movement similarly. Under these assumptions, we treat $\mathbf{d}_c$ as a random variable whose posterior is inferred by incorporating information about the environment occupancy around c, used as prior, and observations coming from trajectory data. Formally, according to assumption (i), given a certain neighborhood $\mathbf{N}_{c_r} \subset \mathbf{M}$ around a reference cell c_r, our first target is to learn the prior $\bar{\mathbf{d}}_{c_r} = \mathsf{P}(\mathbf{d}_{c_r} \mid \mathbf{w}_{c_r})$, where $\mathbf{w}_{c_r} = \{s(c) \mid c \in \mathbf{N}_{c_r}\}$, *i.e.*, a window over the occupancy map describing the geometry of the environment around c_r. We also propose that, according to assumption (ii), given a different environment $\mathbf{M}'$, and a reference cell $c'_r \in \mathbf{M}'$, $\mathbf{w}_{c'_r} \approx \mathbf{w}_{c_r} \implies \bar{\mathbf{d}}_{c'_r} \approx \bar{\mathbf{d}}_{c_r}$.

In Sec. III-A we will present how to obtain the posterior for $\mathbf{d}_c$, while in Sec. III-B we will propose a method to approximate the prior $\bar{\mathbf{d}}_c$ through deep learning.

A. Posterior inference using conjugate prior

Given the prior $\bar{\mathbf{d}}_c$ and a set of observations $\mathbb{D}_c = \{\delta_1, \ldots \delta_N\} \sim \mathrm{Cat}(k, \mathbf{d}_c)$, *e.g.*, obtained from trajectory data of people moving in the environment, we can infer the posterior using the Dirichlet distribution, conjugate prior for the categorical distribution [13].

Let $\alpha > 0$ be a concentration hyperparameter, indicating our trust in the prior, and $\mathbf{q}_c = (q_{1c}, \ldots, q_{kc})$ represent the number of occurrences of direction i in $\mathbb{D}_c$ such that $q_{ic} = \sum_{j=1}^{N} \mathbf{1}_i(\delta_j)$. Then the posterior $\mathsf{P}(\mathbf{d}_c \mid \alpha, \bar{\mathbf{d}}_c, \mathbb{D}_c) \sim \mathrm{Dir}(k, \mathbf{q}_c + \alpha \cdot \bar{\mathbf{d}}_c)$, which allows us to calculate the expected value for each directional probability d_{ic} as:

$$\mathrm{E}[d_{ic} \mid \alpha, \bar{\mathbf{d}}_c, \mathbb{D}_c] = \frac{q_{ic} + \alpha \cdot \bar{d}_{ic}}{N + \alpha} \ . \tag{2}$$

The set of all $d_{ic}\ \forall i \in [1, k]; c \in \mathbf{M}$ defines the complete posterior $\mathbf{d}$. Whenever new observations become available, (2) is also used to updated the belief over the posterior.

B. Parametric approximation of the prior

We learn $g_\theta \approx \mathsf{P}(\mathbf{d} \mid \mathbf{w})$, *i.e.*, a parametric approximation of the prior defined by the parametrization θ, that we model as a FC-DenseNet architecture [14] following previous literature [5], [6]. The network takes as input a 64×64 window over an occupancy grid map, processes it over several densely connected blocks of convolutional layers and max-pooling layers, before up-sampling it through transposed convolutions and outputting the k-dimensional transition probability distribution $\bar{\mathbf{d}}_{c_r}$, with c_r being the center pixel of the input window. In this study we use $k = 8$ in order to model the probability of moving in the direction

of each of the eight neighboring cells to c_r. Please refer to [14] for the exact composition of the dense blocks.

One thing to note is that most occupancy grid maps are built at very high resolution (usually 0.05 m to 0.1 m), but for MoDs modelling human traffic, those resolutions are too dense, and they are usually constructed at around 0.4 m to 1 m per cell [2]. Therefore, to be able to build models at arbitrary output grid resolutions, we scale the grid resolution for the input of the network, by interpolating from the original grid resolution of the occupancy map.

IV. Implementation

We train g_θ in a supervised fashion by using a dataset of pairs $(\mathbf{w}_c, \hat{\mathbf{d}}_c)$ of occupancy windows $\mathbf{w}_c$ with their corresponding groundtruth transitions $\hat{\mathbf{d}}_c$. As loss we use mean squared error (MSE) between the predicted transition probabilities and the groundtruth. We train for 120 epochs, using Adam as optimizer with a fixed learning rate of 0.001.

To extend the amount of available data, we augment each input-output pair randomly by vertical and/or horizontal flipping followed by a random rotation of either 0, $\frac{1}{2}\pi$, π, or $\frac{3}{2}\pi\,\mathrm{rad}$, with equal probability. When we perform these augmentations, the groundtruth transition probabilities are transformed accordingly to still match the transformed input window.

The dataset we use for training our model is the *ATC Dataset*, containing real pedestrian data from the ATC mall (The Asia and Pacific Trade Center, Osaka, Japan, first described by [15]). This dataset was collected with a system consisting of multiple 3D range sensors, covering an area of about 900 m^2. From the ATC dataset, we picked Wednesday November 14, 2012 for training and Saturday November 18, 2012 for testing, which we will refer to as *ATC-W* and *ATC-S* respectively. For each day, we built a ground-truth MoD using the floor field algorithm [10] which constructs a per-cell 8-directional transition model by accumulation directly from trajectory data. We refer to these models as $\hat{\mathbf{d}}^W$ and $\hat{\mathbf{d}}^S$ respectively. For training then, we will use a dataset composed of 1479 pairs of cells $(\mathbf{w}_c, \hat{\mathbf{d}}^W_c)$, while a dataset of 1360 pairs $(\mathbf{w}_c, \hat{\mathbf{d}}^S_c)$ will be used for validation. In both, $\mathbf{w}_c$ is a window around cell c extracted from s^A.

As dataset representing an unseen environment, we use the *KTH Track Dataset* [16], which we will refer to as *KTH*. Data from this dataset is never seen during training and is only used for evaluation. In this dataset, 6251 human trajectory data were collected by an RGB-D camera mounted on a Scitos G5 robot navigating through University of Birmingham library. An occupancy grid map of the environment s^K at 0.05 m/pixel resolution is also available with the dataset. Similarly to ATC, for this dataset we learn a gold standard floor field model $\hat{\mathbf{d}}^K$ using all trajectories available for this dataset.

V. Experiments

In order to evaluate the proposed method, we want to assess how informative is the learned prior when modelling human motion within the same environment, but on a different day (*ATC-S*). Moreover, we want to know how well does the Bayesian map of dynamics approach transfer to a different environment (*KTH*).

In these experiments, we want to compare the performance of the proposed transition model against the gold standard model, *i.e.*, the floor field model built only using trajectory data, as well as a Bayesian model using an uninformed uniform prior.

As a metric, we will compute the likelihood for a trajectory dataset to be predicted by each model. Formally, each trajectory dataset $\mathbb{D}$ is a sequence of observations defined by their xy-coordinates and a motion angle $\delta \in [0, 2\pi)$, *i.e.*, $\mathbb{D} = [(x_1, y_1, \delta_1), \ldots, (x_N, y_N, \delta_N)]$ representing the position a person was standing in (in world coordinates) and the direction they were moving towards. Then, given a transition model $\mathbf{d}$, the average likelihood for a dataset $\mathbb{D}$ is computed as $\mathcal{L}(\mathbb{D} \mid \mathbf{d}) = \frac{1}{N}\sum_{j=1}^{N} \mathsf{P}(\delta_j \mid \mathbf{d}_{c_j})$ where $\mathbf{d}_{c_j}$ refers to the transition model for the grid cell c_j containing the coordinates (x_j, y_j), and $\mathsf{P}(\delta_j \mid \mathbf{d}_{c_j})$ is computed following (1).

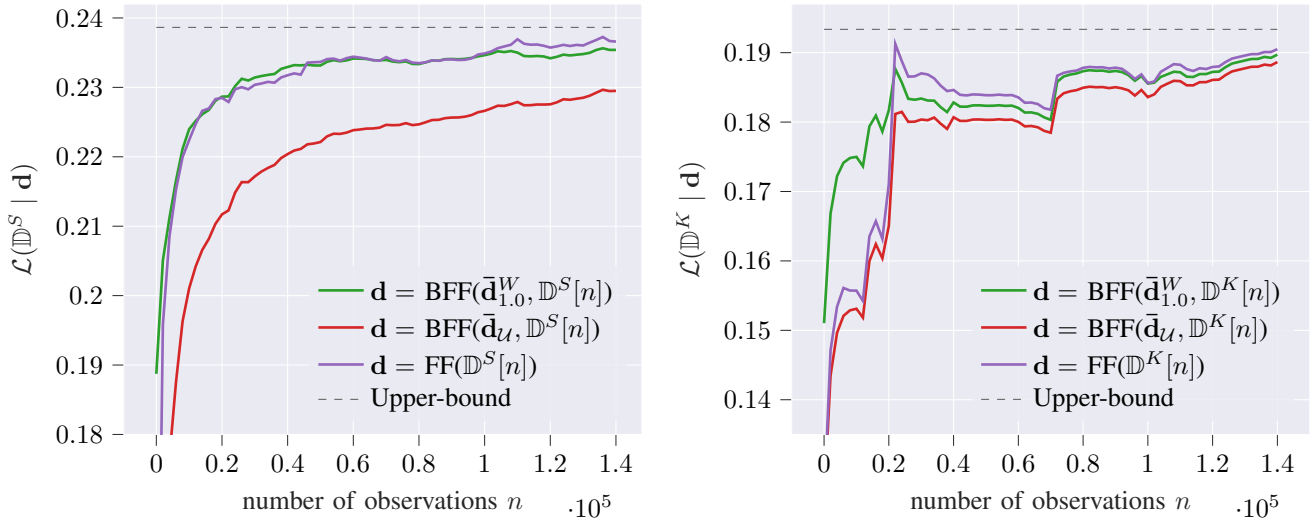
As datasets for our evaluation we use all observations from *ATC-S* and *KTH*, namely $\mathbb{D}^S$ and $\mathbb{D}^K$ respectively. $\mathbb{D}^S$ contains 51 844 trajectories, amounting to 8 533 469 observations in total, while $\mathbb{D}^K$ contains 6251 trajectories, amounting to 421 111 observations in total.

As first step, we train the network-based prior used by our method on *ATC-W* with an input resolution of 1 m/cell. All other training parameters are presented in Sec. IV. We will refer to this prior as $\bar{\mathbf{d}}^W_{1.0}$. Moreover, we will refer to the uniform prior as $\bar{\mathbf{d}}_{\mathcal{U}}$. For both methods utilizing a prior, we use $\alpha = 5$.

We want to measure the performance of each model as a function of the number of observations used to build it. Our expectation is that our method, by relying on the prior, will improve the performance especially under low numbers of observations; as the number of available observation grows, the benefits provided by the prior will diminish and the floor field model will be able to capture the dataset equally well. In order to validate this hypothesis, we split each dataset in chunks of 2000 observations, and then evaluate each method after growing the dataset one chunks at a time. We will use the notation $\mathbb{D}[n]$ to refer to the subset composed of the first n observations of dataset $\mathbb{D}$. We also use the notation BFF($\bar{\mathbf{d}}, \mathbb{D}[n]$) to refer to the posterior of our Bayesian floor field method using prior $\bar{\mathbf{d}}$ and the $\mathbb{D}[n]$ observation dataset, and FF($\mathbb{D}[n]$) to refer to the floor field model using the $\mathbb{D}[n]$ observation dataset.

Fig. 2 shows the performance of each method in each testing environment as a function of n, *i.e.*, the amount of observations available. To improve readability of the plots, we crop the visualization at $n = 140000$; after that point each method has had enough data to converge and performs approximately the same. 140 000 observations correspond to approximately 850 trajectories on *ATC-S* and 2078 trajectories in *KTH*.

One thing to note about the metric used, the likelihood $\mathcal{L}$



(a) *ATC-S* (b) *KTH*

Fig. 2: Performance improvements by the proposed Bayesian Floor Field (BFF) over state-of-the-art Floor Field (FF), measured in function of the amount of observations available. For each graph, the dotted line represents the likelihood upper-bound (best viewed in color).

TABLE I: Average likelihood and percentage of the range between lower and upper bounds at $n = 0$ when using different priors.

	ATC-S		*KTH*	
Prior	$\mathcal{L}$	%	$\mathcal{L}$	%
$\bar{\mathbf{d}}_{\mathcal{U}}$	0.125	0.0	0.125	0.0
$\bar{\mathbf{d}}^W_{1.0}$	**0.189**	**57.7**	**0.151**	**38.2**
Upper-bound	0.236	100.0	0.193	100.0

has an dataset-specific upper-bound, given by the intrinsic ambiguity of behaviors in each dataset (it is impossible to predict with absolute certainty the direction a person in a certain location will move towards). These upper-bounds are reached when using all available trajectories to build the transition model and then measuring the likelihood of those same trajectories given those complete models. These upper-bounds are $\mathcal{L}(\mathbb{D}^S \mid \hat{\mathbf{d}}^S) = 0.2386$ for *ATC-S* and $\mathcal{L}(\mathbb{D}^K \mid \hat{\mathbf{d}}^K) = 0.1933$ for *KTH* respectively, and are shown as a dotted line in Fig. 2.

When considering *ATC-S* (Fig. 2a), *i.e.*, the same environment the network was trained on, but using trajectories from a different day, we immediately can notice that adding an uninformed prior, *i.e.*, $\bar{\mathbf{d}}_{\mathcal{U}}$, considerably hinders the performance. However, we can see that our learned prior only marginally improve the performance of the floor field model.

However, when looking at the generalization capabilities of our proposed method on the unseen *KTH* environment (Fig. 2b), we can make a couple of observations. $\bar{\mathbf{d}}^W_{1.0}$ is actually able to generalize and improves the performance over floor field model up to $n = 20000$ (297 trajectories)

Another interesting observation comes when looking at $n = 0$ in Fig. 2b, *i.e.*, when only the prior is used in a completely unseen environment with no data from that domain. Tab. I presents the results for $n = 0$ more explicitly. It can be seen that our occupancy-based prior is performing better than the uninformed one, with $\bar{\mathbf{d}}^W_{1.0}$ showing the highest likelihood, *i.e.*, $\mathcal{L} = 0.151$, which is around 38% better, when considering the range up to the upper-bound for *KTH*. This is a remarkable result that demonstrates at the fact that knowledge about the relationship between environment occupancy and people flow transfer across environments.

VI. Conclusions

In this work, we presented a novel approach to infer maps of dynamics (MoDs) from architectural geometry and transfer the learned model to new unseen environments. Moreover we proposed a mapping method using Bayesian inference to combine occupancy-based and trajectory-based MoDs. We evaluated the generalization ability of the proposed method on human trajectories in different large-scale environments, showing that, when tasked to predict pedestrian motion across environments, the proposed method is able to improve performance while requiring less trajectories.

In conclusion, both the ability of the proposed method to generalize to unseen large-scale buildings and its ability to combine different type of data is unprecedented in MoDs literature. When considering these findings from a broader perspective, they illuminate an untapped potential in robotic mapping: most mapping approaches tend to focus on one property of the environment at a time, however modeling the latent correlation between different properties can increase data efficiency and provide richer maps. Studying which environment characteristics are good predictors for properties that are time-consuming or expensive to map, like people flow, will be crucial to make complex maps more ubiquitous in robotics.

REFERENCES

[1] T. P. Kucner, M. Magnusson, S. Mghames, L. Palmieri, F. Verdoja, C. S. Swaminathan, T. Krajník, E. Schaffernicht, N. Bellotto, M. Hanheide, and A. J. Lilienthal, "Survey of maps of dynamics for mobile robots," *The International Journal of Robotics Research*, vol. 42, no. 11, pp. 977–1006, 2023.

[2] T. P. Kucner, A. J. Lilienthal, M. Magnusson, L. Palmieri, and C. S. Swaminathan, *Probabilistic Mapping of Spatial Motion Patterns for Mobile Robots*. Springer International Publishing, 2020.

[3] T. Lai, W. Zhi, and F. Ramos, "Occ-traj120: Occupancy maps with associated trajectories," *CoRR*, 2019.

[4] W. Zhi, T. Lai, L. Ott, and F. Ramos, "Trajectory Generation in New Environments from Past Experiences," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sep. 2021, pp. 7911–7918.

[5] J. Doellinger, M. Spies, and W. Burgard, "Predicting Occupancy Distributions of Walking Humans With Convolutional Neural Networks," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1522–1528, Jul. 2018.

[6] J. Doellinger, V. S. Prabhakaran, L. Fu, and M. Spies, "Environment-Aware Multi-Target Tracking of Pedestrians," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1831–1837, Apr. 2019.

[7] R. Senanayake and F. Ramos, "Bayesian hilbert maps for dynamic continuous occupancy mapping," in *Proceedings of the 1st Annual Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, S. Levine, V. Vanhoucke, and K. Goldberg, Eds., vol. 78. PMLR, 2017, pp. 458–471.

[8] S. Molina, G. Cielniak, T. Krajník, and T. Duckett, "Modelling and predicting rhythmic flow patterns in dynamic environments," in *Towards Autonomous Robotic Systems*, M. Giuliani, T. Assaf, and M. E. Giannaccini, Eds. Cham: Springer International Publishing, 2018, pp. 135–146.

[9] J. Shi and T. P. Kucner, "Learning State-Space Models for Mapping Spatial Motion Patterns," in *2023 European Conference on Mobile Robots (ECMR)*, Sep. 2023, pp. 1–6.

[10] C. Burstedde, K. Klauck, A. Schadschneider, and J. Zittartz, "Simulation of pedestrian dynamics using a two-dimensional cellular automaton," *Physica A: Statistical Mechanics and its Applications*, vol. 295, no. 3, pp. 507–525, 2001.

[11] T. P. Kucner, M. Magnusson, E. Schaffernicht, V. H. Bennetts, and A. J. Lilienthal, "Enabling flow awareness for mobile robots in partially observable environments," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 1093–1100, 4 2017.

[12] H. Moravec and A. Elfes, "High resolution maps from wide angle sonar," in *1985 IEEE International Conference on Robotics and Automation Proceedings*, Mar. 1985, pp. 116–121.

[13] T. P. Minka, "Bayesian inference, entropy, and the multinomial distribution," Jan. 2003, tutorial. [Online]. Available: https://tminka.github.io/papers/minka-multinomial.pdf

[14] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation," *arXiv:1611.09326 [cs]*, Oct. 2017.

[15] D. Brščić, T. Kanda, T. Ikeda, and T. Miyashita, "Person tracking in large public spaces using 3-d range sensors," *IEEE Transactions on Human-Machine Systems*, vol. 43, no. 6, pp. 522–534, 2013.

[16] C. Dondrup, N. Bellotto, F. Jovan, and M. Hanheide, "Real-Time Multisensor People Tracking for Human-Robot Spatial Interaction," in *International Conference on Robotics and Automation (ICRA) - Workshop on Machine Learning for Social Robotics*, 2015.