This is the author's version of an article that has been published in the IROS 2022 proceedings.
Changes were made to this version by the publisher prior to publication.
The final version of record is available at https://doi.org/10.1109/IROS47612.2022.9982067

Non-Parametric Modeling of Spatio-Temporal Human Activity Based on Mobile Robot Observations

Marvin Stuede and Moritz Schappler[1]

***Abstract*— This work presents a non-parametric spatio-temporal model for mapping human activity by mobile autonomous robots in a long-term context. Based on Variational Gaussian Process Regression, the model incorporates prior information of spatial and temporal-periodic dependencies to create a continuous representation of human occurrences. The inhomogeneous data distribution resulting from movements of the robot is included in the model via a heteroscedastic likelihood function and can be accounted for as predictive uncertainty. Using a sparse formulation, data sets over multiple weeks and several hundred square meters can be used for model creation. The experimental evaluation, based on multi-week data sets, demonstrates that the proposed approach outperforms the state of the art both in terms of predictive quality and subsequent path planning.**

I. Introduction

The ability to create environmental models is a crucial requirement for the autonomy of mobile robots. Especially in long-term applications, the consideration of environmental dynamics has proven to be useful for localization or navigation purposes [1]. Human behavior represents a major influencing factor on environmental dynamics, particularly for applications in service robotics or autonomous driving. To ensure that robots are accepted by humans and not perceived as a disturbance, they should adapt to human behavior, e.g. in terms of where they move or the timing of their tasks. Accurate models of human activity, i.e. spatio-temporal occurrences and movements of pedestrians, can help robots to achieve this purpose, e.g. by improving navigation [2], task planning [3] or human-centered task execution [4]. As current research shows [5], [6], continuous representations can better reflect human activity than approaches that use spatial or temporal discretization, as interdependencies between data points can be accounted for. Following this idea, we present CoPA-Map (**Co**ntinuous **P**edestrian **A**ctivity Map), a non-parametric model for long-term prediction of human presence. We focus on the use in mobile robotics, which is characterized by varying dwell times at different locations and thus leads to an inhomogeneous, or sparse, distribution of measurement data. The model is implemented using multi-latent Gaussian Process Regression (GPR), allowing time- and location-dependent variances to be incorporated using a heteroscedastic likelihood function. Locations with high variability of observed human activity, for example, due to short dwell times, as well as outliers are thereby given lower weight by adjusting the likelihood variance during hyperparameter optimization. This can also be accounted for in the resulting predictive uncertainty to indicate areas of the input space which require further exploration or data collection (see Fig. 1 for an example). Gaussian Processes (GPs) are also particularly suitable for the given application in that spatial correlations or temporal characteristics can be taken into account as prior information. Based on a data-driven initialization procedure, we therefore create a multi-dimensional kernel that encodes long-term periodic patterns resulting from people's routines. The code of our method is available online [7].

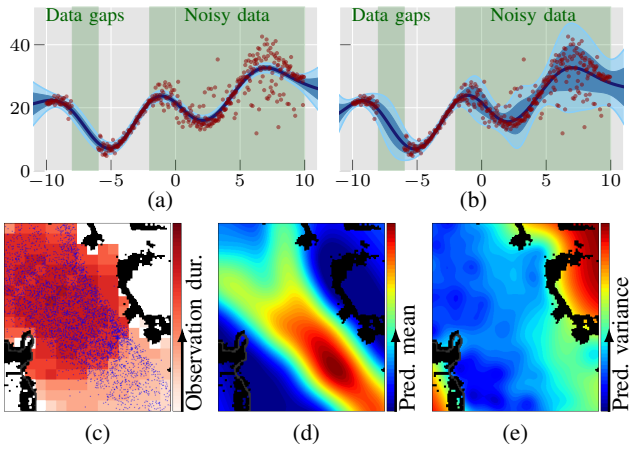


Fig. 1: Upper images: Exemplary 1D dataset with gaps and noise, fitted with a Gaussian process model with Gaussian likelihood (a) and a Gaussian process with a heteroscedastic Gaussian likelihood (b). Mobile robots detect a varying number of pedestrians (blue dots), depending on the observation duration at different locations (c). Our model aims to infer a continuous rate function of human activity, which compensates for these effects (d). Locations with fewer or irregular detections are indicated by the predictive variance (e), based on the heteroscedastic likelihood.

The remainder of this paper is structured as follows: the next section II gives an overview of related work. Sec. III introduces GPR and corresponding preliminaries. Sec. IV presents our method CoPA-Map, which is evaluated in Sec. V and a conclusion is given in Sec. VI.

II. Related work

Approaches to modeling human activity generally consider spatial or temporal variations or a combination of both. For an indication of local variability, many models discretize the spatial coordinates, e.g. using a grid, so that different locations are considered separately. In [8] a directional grid map is presented, that probabilistically models long-term human motion through angular directions. Angular representations,

[1]Authors are with the Leibniz University Hannover, Institute of Mechatronic Systems, D-30823 Garbsen, Germany, marvin.stuede@imes.uni-hannover.de

that also incorporate motion speed and partial observability are presented in [9]. Instead of separating the environment into discrete locations, other approaches create continuous representations for short-term trajectory predictions as in [10] or [11]. In [2] spatially continuous navigational maps by observation of human trajectories are created, with a particular focus on integrating a prior path enabled by a Gaussian Process framework. Apart from the modeling of human activity, spatially continuous models have been successfully used for occupancy mapping of static objects [12]. Later works [13], [14] also incorporate environmental dynamics to create long-term maps of occupancy. These non-parametric methods are typically kernel-based and therefore can distinguish well between empty and occupied space, and can also capture nonlinear or obstructed patterns. However, the aforementioned works [2], [8]–[14] focus on spatial relations and neglect temporal variations, especially with respect to long-term changes. Models which consider long-term temporal patterns usually focus on periodic changes, which can be modeled kernel-based [15] or with spectral analysis, e.g. by the FreMEn method [1]. FreMEn is a method for non-uniform frequency transforms with an application to mobile robotics and was originally developed to model the evolution of binary states over time, such as cells of an occupancy grid. Therefore, extensions have been made to model human activity quantitatively using spatially discrete Poisson processes with respect to intensities [16] or predominant directions of human flow [17]. As these methods either only consider temporal variations [15] or neglect interdependencies of separate spatial regions [16], [17], authors of [5] proposed a spatio-temporal continuous model of human presence. The model is based on a projection of data points to a circular space with subsequent clustering by Gaussian Mixture Models (GMMs) and was later extended to incorporate human flow [6]. However, since clustering is performed directly on the data points (people detections), it is prone to erroneous predictions when the robotic system moves through the environment and collects varying amounts of data at different locations.

In summary, the long-term prediction of human activity, which is suitable for mobile robotic applications, requires further research. The contributions of this paper are therefore: 1) A model for long-term predictions of human presence that compensates for inhomogeneous data distribution resulting from a moving robot and incorporates spatio-temporal interdependencies due to its continuous representation, 2) a data-specific routine for initializing hyperparameters representing periodic changes in human activity which significantly enhances model convergence, 3) experiments of the method on real-world datasets.

III. Preliminaries

A. Gaussian Process Regression (GPR)

For a dataset of n training inputs $\boldsymbol{X} = \left\{\boldsymbol{x}_i \in \mathbb{R}^d\right\}_{i=1}^n$ and observations $\boldsymbol{y} = \{y_i \in \mathbb{R}\}_{i=1}^n$ the standard formulation of GPR aims at inferring a latent function $f : \mathbb{R}^d \to \mathbb{R}$ via a noisy observation model

$$y_i = f(\boldsymbol{x}_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2). \tag{1}$$

The Gaussian Process is defined as a distribution over functions $\mathbf{f} = f(\boldsymbol{x}) \sim \mathcal{GP}(\mu_f(\boldsymbol{x}), k_f(\boldsymbol{x}, \boldsymbol{x}'))$ with mean function $\mu_f(\boldsymbol{x})$ and covariance function $k_f(\boldsymbol{x}, \boldsymbol{x}')$.

B. Prior Approximation via Inducing Inputs

The most prominent weakness of standard GPs is their cubic complexity in the number of training inputs $\mathcal{O}(n^3)$ due to the inversion of the $n \times n$ kernel matrix $\mathbf{K_{ff}} = k_f(\boldsymbol{X}, \boldsymbol{X})$. This limits their usability, especially for applications in robotics and on large datasets. A common approach to overcome this problem is to sparsely approximate the kernel matrix $\mathbf{K_{ff}}$ using the Nyström low-rank representation $\mathbf{K_{ff}} \approx \mathbf{K_{fu_f}} \mathbf{K_{u_f u_f}^{-1}} \mathbf{K_{fu_f}^{\mathrm{T}}}$. Therefore, a number of m inducing points (or pseudo-inputs), where $m \ll n$, must be chosen at locations $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^m$ to optimally represent the training data. The corresponding function values are denoted as $\mathbf{u}_f = f(\mathbf{Z})$. This decreases the computational cost to $\mathcal{O}(m^2 n)$, which can further be reduced by variational approximations utilizing Stochastic Gradient Descent (SGD) (see IV-D). As the quality of the approximation largely depends on the number and location of inducing inputs, it is suitable to treat the inducing points as hyperparameters, and optimize their locations $\mathbf{Z}$ with respect to the marginal likelihood.

C. Variational Inference for Multiple Latent Functions

In the case of heteroscedastic GPR, parameters of the likelihood function can vary with the input. For Gaussian likelihoods this changes the original GP model (eq. 1) to $y_i \sim \mathcal{N}(f(\boldsymbol{x}_i), \zeta(g(\boldsymbol{x}_i)))$, where $g(\boldsymbol{x}) \sim \mathcal{GP}(\mu_g(\boldsymbol{x}), k_g(\boldsymbol{x}, \boldsymbol{x}'))$ is a second latent function that can also be modeled by a GP. The function $\zeta(.) : \mathbb{R}^d \to \mathbb{R}_+^d$ is a link function to guarantee positive values for the noise parameter [18].

In a model with multiple latent functions the marginal likelihood $p(\boldsymbol{y})$ is not analytically tractable and posterior approximations are required. Instead of calculating the intractable posterior $p(\mathbf{f}, \mathbf{g}|\boldsymbol{y})$, it can be lower bounded with variational distributions $q(\mathbf{f})$ and $q(\mathbf{g})$, a technique called *variational inference*. The main principle of this technique is the estimation of the parameters of $q(\mathbf{f})$ and $q(\mathbf{g})$ by minimizing their distance to the true posterior distribution measured by the Kullback-Leibler-divergence $\mathrm{KL}\left(q(\mathbf{f})q(\mathbf{g}) \| p(\mathbf{f}, \mathbf{g}|\boldsymbol{y})\right)$. Assuming that the latent functions $\mathbf{f}$ and $\mathbf{g}$ are a priori independent for each data point, Saul et al. [19] derive the variational lower bound

$$\begin{aligned}\mathcal{L} = &\sum_{i=1}^{n} \int q(\mathbf{f}_i)\, q(\mathbf{g}_i) \log p(\boldsymbol{y}_i \mid \mathbf{f}_i, \mathbf{g}_i)\, \mathrm{d}\mathbf{f}_i\, \mathrm{d}\mathbf{g}_i \\ &- \mathrm{KL}\left(q(\mathbf{u}_f) \| p(\mathbf{u}_f)\right) - \mathrm{KL}\left(q(\mathbf{u}_g) \| p(\mathbf{u}_g)\right).\end{aligned} \tag{2}$$

This bound leverages the aforementioned sparse formulation and aims at calculating sparse approximate posteriors as normal distributions $q(\mathbf{u}_f) = \mathcal{N}\left(\mathbf{u}_f | \boldsymbol{\mu}_f, \boldsymbol{S}_f\right)$ and $q(\mathbf{u}_g) = \mathcal{N}\left(\mathbf{u}_g | \boldsymbol{\mu}_g, \boldsymbol{S}_g\right)$ over inducing functions $\mathbf{u}_f$ and $\mathbf{u}_g$. For

$q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mathbf{m}_f, \mathbf{\Sigma}_f)$ follows

$$\mathbf{m}_f = \mathbf{K}_{\mathbf{f}\mathbf{u}_f}\mathbf{K}_{\mathbf{u}_f\mathbf{u}_f}^{-1}\boldsymbol{\mu}_f, \tag{3}$$

$$\mathbf{\Sigma}_f = \mathbf{K}_{\mathbf{f}\mathbf{f}} + \mathbf{K}_{\mathbf{f}\mathbf{u}_f}\mathbf{K}_{\mathbf{u}_f\mathbf{u}_f}^{-1}(\boldsymbol{S}_f - \mathbf{K}_{\mathbf{u}_f\mathbf{u}_f})\mathbf{K}_{\mathbf{u}_f\mathbf{u}_f}^{-1}\mathbf{K}_{\mathbf{u}_f\mathbf{f}}. \tag{4}$$

The equations for $q(\mathbf{g}) = \mathcal{N}(\mathbf{g}|\mathbf{m}_g, \mathbf{\Sigma}_g)$ follow accordingly. Training the model is then realized by minimizing $-\mathcal{L}$ with respect to the variational parameters $\boldsymbol{\mu}_{f,g}$ and $\boldsymbol{S}_{f,g}$ as well as the hyperparameters in the covariance matrices $\mathbf{K}_{**}$. The latter follow from problem-specific covariance functions, which are chosen to account for prior information (see IV-B).

IV. Methods

To form our model, we consider a mobile robot acting in an environment with a known map, sufficiently accurate self-localization within this environment and a sensor for people detection. A detected pedestrian is represented as a 2D-point $\boldsymbol{p}_k = (x_{1,k}, x_{2,k}, t_k)^{\mathrm{T}}$ in world coordinates corresponding to a measurement taken at time t_k. The goal then is to model human activity as an intensity function of space and time, by first defining a count of people c_i within a spatio-temporal domain $\mathcal{S}_i \subset \mathbb{R}^3$ so that $c_i = |\{\boldsymbol{p}_k \in \mathcal{S}_i\}|$. By partitioning the environment into an evenly spaced grid of n cells, we create each domain $\mathcal{S}_i$ as a cell with square spatial shape with edge length r_{s} and temporal resolution τ. Since the robot is moving through the environment, each cell is visible to the robot for a different time period. This time period is calculated based on the field of view (FOV) of the sensor, which can be approximated by a geometrical shape. For example, the projected 2D-detection area of a 3D-Lidar-based detector can be approximated by a circle, which is pruned at known obstacles in the environmental map based on a ray casting model. A people count $c_i \geq 0$ and observation duration $0 < \Delta_i \leq \tau$ is then assigned to each visible cell. Consequently, the robot's deployments over time generate the set of input data $\boldsymbol{X} = \{(x_{1,i}, x_{2,i}, t_i)\}_{i=0}^{n}$, which consists of the spatio-temporal centers of the cells. The corresponding target are the observed rates $\boldsymbol{y} = \{{}^{c_i}/_{\Delta_i}\}_{i=0}^{n}$ of people in each cell. Considering the rates instead of counts is based on the following idea: Since a target value y_i can both be large due to a large c_i or a small Δ_i, it varies more smoothly at edges between areas with shorter and larger observation periods Δ_i. As people move through the environment in a continuous fashion, areas with consistent values y_i then indicate homogeneous activity which merits greater weighting when optimizing the marginal likelihood. However, irregular spatial patterns of the values in $\boldsymbol{y}$ indicate either short observation durations or irregular occurrences of people, which in contrast should be captured by a larger input noise in the likelihood function. In Fig. 2, an overview of the input data and resulting rate $\boldsymbol{y}$ is given, along with a ground truth which was created without any constraints on the FOV or observation duration.

A. Likelihood Function

To fit a model in the GP-framework, a likelihood function must be chosen that best represents the distribution of observations $\boldsymbol{y}$. Count data, such as person occurrences, can e.g. be viewed as events from an inhomogeneous Poisson process [16]. However, this requires strong assumptions on the independence of events (e.g. people cannot arrive in groups), considers discrete data instead of a continuous rate y_i and the variance of the Poisson distribution is directly coupled to its rate parameter. Instead, we consider the rate $y_i \sim \mathcal{N}(f(\boldsymbol{x}_i), \sigma_i^2)$, to be normally distributed with input-dependent noise σ_i^2, which can be defined independently from the latent mean function $f(\boldsymbol{x}_i)$ and makes training less prone to outliers. When standardizing the target values y_i to zero-mean and a standard deviation of one, this consistently leads to better results than strictly positive likelihoods, such as the Gamma distribution. The latter would additionally require manually tuned normalization for different input datasets to achieve consistent results. As the variance is defined by a latent function $\sigma_i^2 = \zeta\left(g(\boldsymbol{x}_i)\right)$, we chose the softplus function as link function ζ to ensure for positive values. Although the latent function $f(\boldsymbol{x}_i)$ can result in negative values, the rescaled predictive output of a tuned model contained very few zero-crossings on all tested datasets, making it sufficient to use the absolute value of the model output for predictions.

B. Definition of Covariance Functions

Covariance functions allow encoding prior beliefs about the latent function of interest and can be viewed as a measure of how *similar* two functions are. Different suitable covariance functions can also be connected as compositions. For the present use case of representing human activity, each data point is separated into its spatial component $\boldsymbol{x}_{\mathrm{s}} \in \mathbb{R}^2$ and temporal component $x_{\mathrm{t}} \in \mathbb{R}$ and the following covariance function is defined

$$k_f\left(\boldsymbol{x}_{\mathrm{s}}, x_{\mathrm{t}}, \boldsymbol{x}'_{\mathrm{s}}, x'_{\mathrm{t}}\right) = k_{\mathrm{s}}\left(\|\boldsymbol{x}_{\mathrm{s}} - \boldsymbol{x}'_{\mathrm{s}}\|_2\right) k_{\mathrm{t}}\left(|x_{\mathrm{t}} - x'_{\mathrm{t}}|\right). \tag{5}$$

This multidimensional product kernel connects a spatial component k_{s} with a temporal component k_{t} and results in a prior over functions that varies across all three dimensions. As the spatial kernel, the Matérn-$5/2$ covariance function

$$k_{\mathrm{s}}(r) = \sigma_{\mathrm{s}}^2\left(1 + \frac{\sqrt{5}r}{l_{\mathrm{s}}} + \frac{5r^2}{3l_{\mathrm{s}}^2}\right)\exp\left(-\frac{\sqrt{5}r}{l_{\mathrm{s}}}\right) \tag{6}$$

is chosen, where l_{s} and σ_{s}^2 are hyperparameters. This type of covariance function is a common choice to model structural correlations, as it provides a good balance between smoothness and capturing sudden changes [20].

Oftentimes, human activity can be considered periodic in time. The number of people at different locations is subject to a regularity that is determined, for example, by the time of day, working hours or store opening hours. Therefore, as prior information for the time-dependent person rate, we specify the rate to be subject to periodicities. This can be encoded by a periodic kernel [21], which is defined as a sum of trigonometric functions

$$k_{\mathrm{t}}(r) = \sum_{i=0}^{\psi}\sigma_{\mathrm{t},i}^2\exp\left(-\frac{1}{2}\frac{\sin^2\left(\gamma_i^{-1}r\right)}{l_{\mathrm{t},i}^2}\right) \tag{7}$$

where the variances $\sigma^2_{\mathrm{t},i}$, periods γ_i and lengthscales $l_{\mathrm{t},i}$ are hyperparameters. The variances $\sigma^2_{\mathrm{t},i}$ determine the overall influence of the specific component and $l_{\mathrm{t},i}$ controls the smoothness.

Regarding human activity, the kernel k_f thus represents two important properties: 1.) Spatial continuity, i.e. if people are seen at a specific location it is more likely to also see people at locations that are very close. Since humans move through space in a continuous manner, this property is desirable to model. 2.) Temporal periodicity, i.e. when people are seen repeatedly at a specific location (e.g. every morning at an entrance) it is likely to see people there in the future at that specific point in time. The kernel k_g corresponding to the latent function $\mathbf{g}$ is simply realized by a radial basis function (RBF) kernel. This is sufficient since then the predictive variances of different areas align for larger prediction horizons.

C. Initialization of Hyperparameters

Due to the dependence on many data points as well as hyperparameters, optimization of the lower bound (eq. 2) is prone to get stuck in local minima. A major influencing factor is the initial guess of the hyperparameters. In the present scenario, this applies in particular to the periods γ_i and variances $\sigma^2_{\mathrm{t},i}$ of the temporal kernel k_{t} (eq. 7). With algorithm 1, we therefore propose a method to obtain the characteristic temporal periods of a spatial domain based on non-uniform frequency analysis and a subsequent clustering step. The algorithm builds on the idea [1] of transferring the time-dependent activities at different locations into the frequency spectrum and making an approximation via a Fourier series with a reduced number of components. By squashing the cells $\mathcal{S}_i$ of the spatio-temporal grid along the temporal dimension, a spatial grid with a time series of rates $\boldsymbol{y}_s \subset \boldsymbol{y}$ for each spatial cell s results. A subset $\mathcal{T}$, containing a number of l spatial cells, is then taken from this spatial grid by sampling, where each cell is given a weight of its total counts over all timesteps. This results in the selection of cells that are more likely to have high activity but does not completely exclude cells with lower activity. Due to the movement of the robot, the rates within $\boldsymbol{y}_s$ are non-equidistant with respect to the time of their detection. The conversion to the frequency domain is therefore made by means of the Non-uniform discrete Fourier transform (NUDFT) [22] (line 5). This requires a set of candidate periods O, which is defined to contain equally spaced periods within an interval (e.g. between one hour and seven days). Additionally, the algorithm needs an upper limit $\psi_{\max}$ of periods to check and a scaling factor $\sigma^2_{\max}$ as the maximum variance. The optimal number of periods for each cell is determined by five-fold cross-validation, by comparing the test data with the signal that was reconstructed from a reduced number of frequency components (lines 3 to 11). The total number of periods ψ of the whole domain is then calculated as the mean of the number of periods of the cells in $\mathcal{T}$ (line 16). The periods are calculated by weighted k-means clustering, where the complex magnitudes serve as weights (line 17). This ensures that locations with a large recurring number of people are more influential than cells that have less activity.

Algorithm 1: Init. hyperparameters of periodic kernel

```
   Input  : y, O, T, ψ_max, σ²_max
   Output: ψ, γ̂_{1..ψ}, σ̂²_{1..ψ}
1  foreach s ∈ T do
2      Let y_s be the rates at times t_s of spatial cell s;
3      Repeat lines 4 – 11 as cross validation for i = 1..5;
4      Split y_s into contiguous train/test sets y_s^tr / y_s^ts;
5      ξ_i ← NUDFT(t_s^tr, y_s^tr, O);   // To cplx. components
6      for p = 0 to ψ_max do
7          ξ_{i,p} ← p largest complex numbers in ξ_i w.r.t.
             magnitude;
8          O_{i,p} ← Set of periods, corresponding to ξ_{i,p};
9          ŷ_p ← InverseDFT(ξ_{i,p}, O_{i,p});
10         e_{i,p} ← RMSE(ŷ_p, y_s^ts);
11     end
12     i_s, p_s ← arg min_{i,p}(e_{1,0}, ..., e_{5,ψ_max}) ;
13     A_s ← Save element-wise magnitudes of ξ_{i_s,p_s};
14     O_s ← Set of periods, corresponding to ξ_{i_s,p_s};
15 end
16 ψ ← ⌊Mean({p_1, ..., p_|T|})⌋;
17 γ̂_{1..ψ} ← obtain k-means centroids with k = ψ using
     {O_s | s ∈ T} with weights A_s;
18 σ̂²_{1...ψ} ← Sum weights A_s in clusters and normalize to
     [0, σ²_max];
```

As the full covariance matrix $\mathbf{K_{ff}}$ is not computed, but approximated by covariances over inducing points, their positioning is an additional factor influencing the model quality. Although the inducing points are treated as hyperparameters and therefore modified during optimization, proper initialization reduces the time to find sufficient solutions. Given a ratio $\alpha \in (0, 1]$, the number of inducing points is selected as $m = \lfloor \alpha n \rfloor$. The location is then determined via k-means clustering ($k = m$) of the spatio-temporal training inputs $\boldsymbol{X}$, where each input point is weighted by its individual observation time Δ_i. By weighting the inputs, the initial inducing points $\hat{\mathbf{Z}}$ resulting from the algorithm are primarily placed at locations that have been observed for longer periods and hence provide more reliable data. An exemplary arrangement of inducing points is shown in Fig. 2 (b).

D. Model Optimization

As indicated in section III-B, the optimization of the hyperparameters does not scale cubically when latent inducing locations $\mathbf{Z}$ are used. When the covariance matrices of the variational distributions are parametrized as Cholesky $\boldsymbol{S}_f = L_{\mathbf{f}} L_{\mathbf{f}}^{\mathrm{T}}$ and $\boldsymbol{S}_g = L_{\mathbf{g}} L_{\mathbf{g}}^{\mathrm{T}}$, optimizing the lower bound $\mathcal{L}$ (eq. 2) scales with $\mathcal{O}(nm^2 + 2nm)$ [19]. By choosing a ratio parameter of α so that $m \ll n$, model inference is significantly sped up compared to the standard GPR case. We empirically chose a parameter of $\alpha = 0.02$ to obtain a good balance between computational speed and prediction quality on the evaluated datasets. Model optimization is executed for three types of parameters: 1.) Variational parameters corresponding to $q(\mathbf{u}_f) = \mathcal{N}(\mathbf{u}_f | \boldsymbol{\mu}_f, \boldsymbol{S}_f)$ and $q(\mathbf{u}_g) =$

This is the author's version of an article that has been published in the IROS 2022 proceedings.
Changes were made to this version by the publisher prior to publication.
The final version of record is available at https://doi.org/10.1109/IROS47612.2022.9982067

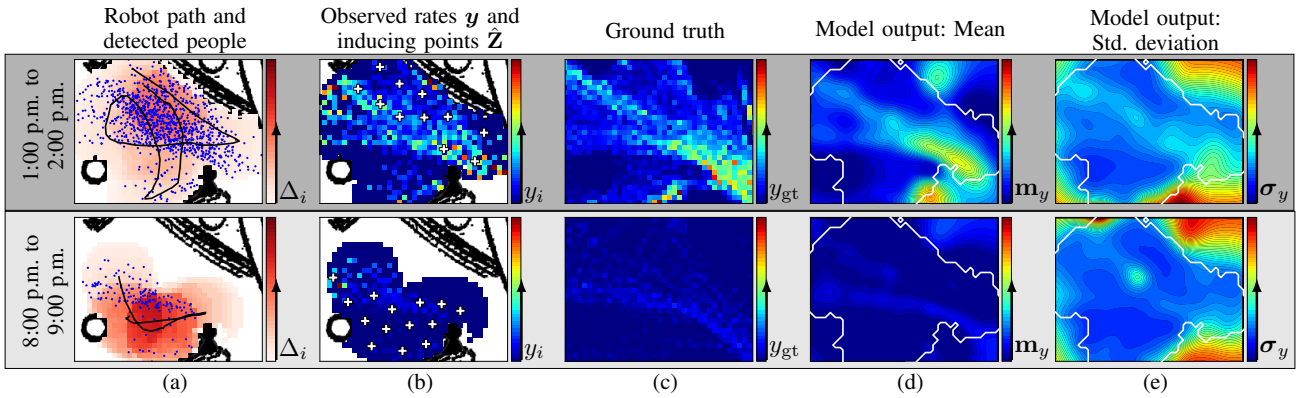


Fig. 2: Input and output data (ATC dataset) of the model at midday (top) and evening (bottom). People detections $\boldsymbol{p}_k$ (blue) and observation durations Δ_i are shown in (a). Resulting rates $\boldsymbol{y}$ for a bin duration of $\tau = 60\,\text{min}$ and initial inducing points (white plusses) are shown in (b). A ground truth with fully observed cells is given for reference in (c). The resulting model outputs are shown in (d) and (e) with a white border indicating the areas that were never observed.

$\mathcal{N}\left(\mathbf{u}_g|\boldsymbol{\mu}_g, \boldsymbol{S}_g\right)$, 2.) the lengthscale, variance and periodicity hyperparameters of the covariance functions k_f and k_g, and 3.) the location of inducing inputs $\mathbf{Z}$. For this, two separate optimization techniques based on SGD are utilized. The variational parameters are optimized using the natural gradient method since the inherent minimization of KL-divergence as part of this method integrates well with the variational framework and leads to fast convergence [23]. The kernel hyperparameters and inducing point locations are optimized with the Adam optimizer [24]. Steps of both optimizers are executed in an alternating fashion, with a linearly decaying learning rate for the first 100 optimization steps. The n integrals as part of the lower bound $\mathcal{L}$ are solved by two-dimensional Gaussian quadratures. As the methods utilize SGD, the optimization can efficiently be separated into mini-batches.

E. Predicting with the Model

After maximization of the variational lower bound and optimization of hyperparameters, the model can be queried via its predictive distribution. For arbitrary new data inputs $\boldsymbol{X}^* = \{\boldsymbol{x}_i^*\}_{i=1}^{n^*}$ the predictive distribution is given as $\int p\left(\boldsymbol{y}_i^* \mid \mathbf{f}_i^*, \mathbf{g}_i^*\right) q\left(\mathbf{f}_i^*\right) q\left(\mathbf{g}_i^*\right) \mathrm{d}\mathbf{f}_i^* \,\mathrm{d}\mathbf{g}_i^*$. This analytically intractable integral can be computed using Gauss-Hermite quadrature to obtain the predictive mean $\mathbf{m}_y$ and variance $\boldsymbol{\sigma}_y^2$. The specific values of the predictive mean depend on the chosen spatial resolution r_s and temporal resolution τ of the input grid. For a subset $\boldsymbol{X}' \subset \boldsymbol{X}^*$ of finite extend (e.g. the FOV of the robot and a given duration) the expected number of people can then be calculated as a point estimate $\frac{1}{r_\text{s}^2 \tau} \int \mathbf{m}_y \,\mathrm{d}\boldsymbol{X}'$. Exemplary model outputs of the mean $\mathbf{m}_y$ and standard deviation $\boldsymbol{\sigma}_y$ for two points in time are shown in Fig. 2 (d) and (e). The uncertainty increases both outside the visited area and in locations where high variability of human activity occurs. In addition to the predictive uncertainty, this indicates in which areas further model exploration could be useful.

V. Experiments

All the following experiments were performed with the same parameterization: $l = 10$, $\psi_\text{max} = 10$, $\sigma^2_\text{max} = 0.95$, $\alpha = 0.02$. The initialization routine (Algorithm 1) was done with a fixed grid resolution of $5.0\,\text{m} \times 60\,\text{min}$, whereas the grid resolution resulting in $\boldsymbol{X}$ was varied for different experiments (respectively specified). The method is implemented in Python based on the GPflow library [25] to perform the training and inference GPU-based.

A. Datasets

We evaluated the model on two freely available long-term datasets containing real-world pedestrian detections. Both datasets represent typical settings for mobile robots but vary in terms of human activity and the number of pedestrians.

ATC Dataset [26]: This dataset contains measurements of tracked pedestrians in a shopping center in Osaka, Japan, covering an area of ca. $900\,\text{m}^2$. Data collection was done with multiple 3D range sensors, every week on Wednesdays and Sundays, resulting in 92 days in total. We downsampled the data to a detection rate of $0.5\,\text{Hz}$, resulting in an average of about 1700 entries per square meter and day. For evaluation, we used a subset of 10 Wednesdays for training and 4 days for testing.

Office Dataset [27]: The second dataset contains tracks of people based on measurements by a single stationary 3D-Lidar in an office environment of the University of Lincoln, England, covering an area of ca. $85\,\text{m}^2$ with averagely about 300 entries per square meter and day. The dataset covers 22 consecutive days, of which we used 10 weekdays for training and 5 weekdays for testing.

As both datasets contain measurements taken by stationary sensors, data collection by a moving robotic system must be simulated. For this purpose, robot trajectories with an average moving speed of $0.5\,\text{m}\,\text{s}^{-1}$ and intermediate stationary stops were specified manually. Then, only the measurements within the FOV of the robot are processed. The FOV is defined by a circle with a fixed radius and is pruned based on the known occupancy maps of the environments to filter

out pedestrians that would be obstructed by static obstacles. Exemplary sets of measurements and robotic paths are shown in Fig. 2 (a).

B. Evaluation Metrics and Baselines

The predictive quality of the model is measured with three criteria. As the evaluation is conducted based on multiple paths, each with different length and area coverage, the first criterion is normalized root mean square error (NRMSE) between model predictions $\hat{y}_i$ and ground truth $y_{\mathrm{gt},i}$

$$\mathrm{NRMSE} = \sqrt{\frac{1}{\bar{y}_{\mathrm{gt}}^2\, n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \left(\hat{y}_i - y_{\mathrm{gt},i}\right)^2}, \tag{8}$$

normalized by the mean test data value $\bar{y}_{\mathrm{gt}}$. The ground truth value is obtained in a similar way as the creation of training data $\boldsymbol{y}$, but ignoring the occupancy map and setting the observation durations $\Delta_i = \tau$. It therefore represents the people count $c_{\mathrm{gt},i}$ during testing time in the cells of the spatio-temporal grid that were visited during training, using the same spatial and temporal resolution. The second criterion is the Chi-square distance

$$\chi^2\mathrm{-distance} = \sum_{i=1}^{n_{\mathrm{test}}} \frac{\left(\hat{y}_i - y_{\mathrm{gt},i}\right)^2}{\left(\hat{y}_i + y_{\mathrm{gt},i}\right)}, \tag{9}$$

where larger values indicate less accurate prediction compared with the test data. These two metrics can be regarded as the standard metrics when comparing human activity or flow models [3], [6], [16], [27]. However, these criteria have limited expressiveness in terms of the *usefulness* of the model e.g. for supporting unobstructed navigation or task planning. Vintr et al. [3] therefore proposed new criteria to evaluate these models based on their ability to support human-aware navigation. The benchmark's main idea is to rate models better which avoid disturbance of people by executing movements of the robot outside of their immediate walking paths. The criterion considers a number of p imaginary navigation scenarios, where a robot should navigate between a set of goal locations at different points in time. The navigational path is planned based on the output of the respective model, where higher activity corresponds to higher path costs. All resulting paths are then ordered ascending by their total cost and the *service disturbance*

$$E(\lfloor pr \rfloor) = \sum_{k=1}^{\lfloor pr \rfloor} e_k \tag{10}$$

is defined as a sum of robot-human encounters e_k during test time. Robot-human encounters e_k are the person detections that occur within a $1\,\mathrm{m}$ radius to the robot while it is simulatively traveling the path at a speed of $0.5\,\mathrm{m\,s^{-1}}$. The value $r \in [0, 1]$ is referred to as *servicing ratio* and defines the number of navigation actions that should be performed. A lower servicing ratio gives more freedom to the robot to discard paths that have high costs, e.g. when the number of expected people is large.

Besides CoPA-Map, the following methods are compared in the evaluation.

The *Maximum-Likelihood* (ML) model calculates the mean of all observed rates in each cell. As a result, the rates are assumed to be constant over time.

Poisson spectral model [16] (Fr-AAM) is a state-of-the-art approach, modeling human activity as an inhomogeneous Poisson process by a spatial grid with a temporally continuous rate function. For each cell of the grid, a spectral analysis based on the FreMEn method [1] is performed repeatedly to obtain the most influential spectral components, from which the predictive signal is then reconstructed.

Warped-Hypertime [6] (WHyTe) is a state-of-the-art approach for continuous activity and flow modeling. It is based on a frequency analysis by the FreMEn method and subsequent projection into a circular space. As training data, it directly uses people detections $\boldsymbol{p}_k$ and outputs the probability of occurrence given an input point. Because of this, we do not directly compare the model output to the quantitative value $c_{\mathrm{gt},i}$, but only include this method in the evaluation of service disturbance. As the method uses a pre-defined number of clusters, we separately trained models with up to seven clusters and only include the variant with the best result. The method is also capable of estimating movement direction and speed, although this is not used in the present evaluation to ensure direct comparability to the other models.

The *Gaussian Process model* (GP-Hom) is based on the same parameters as the proposed method but realized as a Log Gaussian Cox process. The method uses a homoscedastic Poisson likelihood for inference by using a single latent function, transformed with an exponential function to only output positive values which is required for the rate parameter of the Poisson distribution.

C. Validating Hyperparameter Initialization

The first experiment demonstrates the importance of proper initialization of the hyperparameters of the temporal kernel k_{t}. Given the data from an $185\,\mathrm{m}^2$ area of the ATC dataset, the model was trained with different periodic kernels and an RBF-kernel for comparison. Besides our proposed initialization procedure (Alg. 1), we used ten different periodic kernels with variances chosen uniformly randomly in $(0, 1)$ and one to two random periods as multiples of 30 minutes and smaller than 30 hours. Our initialization procedure results in two periods of 12 and 6 hours with variances of 0.9 and 0.42 respectively. Figure 3 shows the negative log-likelihood (NLL) loss during training and the RMSE relative to the ML model on the four independent test days. Due to the high variance of the training data, NLL shows little variation for the periodicity parameter. However, suitable parameters of the periodicities lead to significantly better extrapolations, which is reflected in the values of the RMSE. A complete disregard of periodicities (RBF kernel) further results in unsatisfactory predictive results, since long-term changes cannot be captured and the predictive horizon is limited by the kernel's lengthscale parameter.

This is the author's version of an article that has been published in the IROS 2022 proceedings.
Changes were made to this version by the publisher prior to publication.
The final version of record is available at https://doi.org/10.1109/IROS47612.2022.9982067

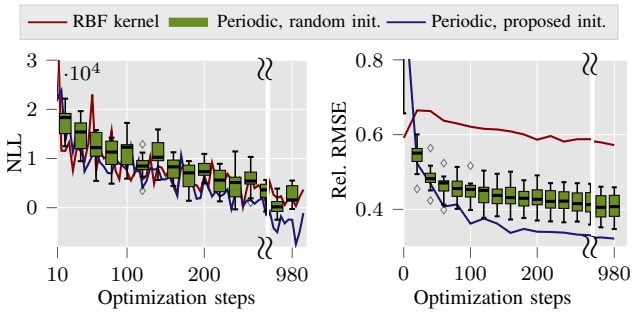


Fig. 3: NLL and relative RMSE (lower is better) for an RBF kernel, a periodic kernel with the proposed initialization routine and periodic kernels with randomly initialized parameters. Both the ML and CoPA-Map model were trained with a 0.5 m × 60 min resolution.

D. Spatio-Temporal Prediction

In order to evaluate the predictive quality of CoPA-Map, we consider the two different scenarios of a *static* and *moving* robot. In the first, a permanently motionless robot is assumed, and a total of five different positions of the robot are considered separately leading to a constant observation time of $\Delta_i = \tau$ for every cell. For the *moving* case, we created 9 different paths with varying spatial coverage of the robot's FOV (between $40\,\mathrm{m}^2$–$70\,\mathrm{m}^2$ for Office and $100\,\mathrm{m}^2$–$200\,\mathrm{m}^2$ for ATC datasets) and different waiting times along the paths. Thus, different cases are covered, where some locations may be permanently in the robot's FOV and others may be visited only a few times a day. Table I shows the results for NRMSE and χ^2-distance. Since these metrics depend on the chosen spatial and temporal resolution (r_s and τ), four different combinations are shown. For the *static* case of the Office dataset, CoPA-Map generally leads to better results than the comparative methods. In the ATC experiments, there is a location in the static case with many people staying for a long time in a small area. Such phenomena can partly be better represented by the discrete models or the homogeneous GP. The advantage of heteroscedastic modeling of CoPA-Map becomes clear in the *moving* case, where the method gives significantly better results. Singularly occurring high target values (e.g. due to very short observation durations) are given less weight by CoPA-Map by adjusting the variance during training. Fr-AAM, on the other hand, strongly approximates areas with high numbers of people, but as a discrete model suffers in terms of error metrics when people appear in slightly different locations in the test data.

The path with the largest area coverage at the smallest resolution (ATC, 0.5 m × 30 min) resulted in 147,500 input points and ca. 2860 inducing points. Training took a maximum of 28 minutes to converge in this case (*Nvidia* GTX1070, i7-8700 CPU, 16 GB RAM). A duration of this magnitude thus makes it possible to repeat the training periodically (e.g. during the charging process) with current data.

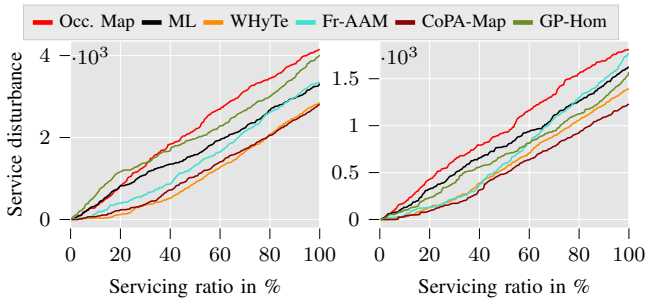
E. Service Disturbance

The ability of the model to provide for active avoidance of areas with high human activity during navigation is also considered in two scenarios. Both the *hallway* and *shops* scenarios are created based on the ATC dataset, the former involving a strong flow of people and the latter a splitting of people movement and longer stationary stays. The input data was again created by a simulated robot movement and the edge weights of the resulting cost map are not directional. Navigation scenarios are created between four positions (A→B→C→D, depicted in Fig. 5), five times per hour between 9 a.m. and 9 p.m. leading to $p = 240$ scenarios for all four test days. As a baseline, the *Occupancy Map* model indicates how many encounters would occur, if the robot would always drive the shortest route by metric distance. Fig. 4 shows the number of encounters (service disturbance) over the servicing ratio r and Fig. 5 gives exemplary model outputs and navigational paths. CoPA-Map and WHyTe as spatially continuous models capture the modality of human activity in the *hallway* scenario significantly better than the discrete models. These models, such as Fr-AAM, often lead to sinuous paths, increasing the number of human encounters. CoPA-Map and WHyTe perform well for smaller service ratios ($< 40\,\%$), since only the navigation tasks in the morning and evening hours are carried out, during which fewer people are expected. Compared to WHyTe, CoPA-Map has advantages when multimodal pedestrian movements occur, as can be seen in the *shops* scenario. As WHyTe does incorporate the detections $\boldsymbol{p}_k$ directly, areas with shorter detection times, and thus fewer detections, might be underrepresented in the data. The underlying GMM is then more likely to underfit. In contrast, e.g. in Fig. 5 (lower) CoPA-Map more accurately represents areas with many pedestrians, resulting in better paths for people avoidance. For a servicing

TABLE I: Predictive performance of the evaluated models for a static and moving robot. NRMSE is given as mean and χ^2-distance as a sum over the results from different paths/locations. χ^2-distance is given as multipliers of 10^4 for brevity of notation.

		Office								ATC							
		0.5 m×30 min		0.5 m×60 min		0.75 m×30 min		0.75 m×60 min		0.5 m×30 min		0.5 m×60 min		0.75 m×30 min		0.75 m×60 min	
		NRMSE	χ^2dst	NRMSE	χ^2dst	NRMSE	χ^2dst	NRMSE	χ^2dst	NRMSE	χ^2dst	NRMSE	χ^2dst	NRMSE	χ^2dst	NRMSE	χ^2dst
Static	ML	2.66	8.80	2.34	8.01	**2.63**	8.66	**2.25**	7.73	1.71	415.33	1.76	**504.95**	1.73	515.34	1.7	517.45
	Fr-AAM	2.71	8.57	2.4	7.82	2.76	9.12	2.33	8.07	**1.71**	**412.19**	1.9	506.18	1.75	**512.83**	1.79	**516.97**
	GP-Hom	2.75	8.05	2.48	8.88	2.79	8.69	2.42	9.02	2.17	528.62	1.7	524.94	**1.68**	516.96	**1.66**	525.44
	CoPA-Map	**2.59**	**6.74**	**2.33**	**6.50**	2.65	**6.95**	2.32	**6.94**	1.78	524.29	**1.69**	578.23	1.8	655.80	1.7	603.43
Moving	ML	2.61	35.27	2.34	32.03	2.57	35.75	2.77	38.89	1.79	456.92	1.94	460.32	1.84	477.26	2.0	481.37
	Fr-AAM	2.6	33.72	2.38	30.76	2.61	33.89	2.67	34.89	**1.36**	**345.00**	1.31	438.58	1.32	420.39	1.28	429.97
	GP-Hom	**2.58**	**31.35**	2.34	33.49	2.56	33.60	2.32	34.74	1.57	649.01	1.5	639.74	1.41	567.46	1.37	576.19
	CoPA-Map	2.61	50.52	**2.13**	**24.91**	**2.33**	**25.28**	**2.19**	**25.20**	1.44	606.58	**0.82**	**172.07**	**0.8**	**164.56**	**0.69**	**127.15**

This is the author's version of an article that has been published in the IROS 2022 proceedings.
Changes were made to this version by the publisher prior to publication.
The final version of record is available at https://doi.org/10.1109/IROS47612.2022.9982067



(a) *Hallway* scenario (b) *Shops* scenario

Fig. 4: Service disturbance (encounters) for two navigation scenarios on the ATC dataset (lower is better). Smaller servicing ratios give more freedom to avoid peak hours of human activity and indicate if a model accurately captures temporal variations.

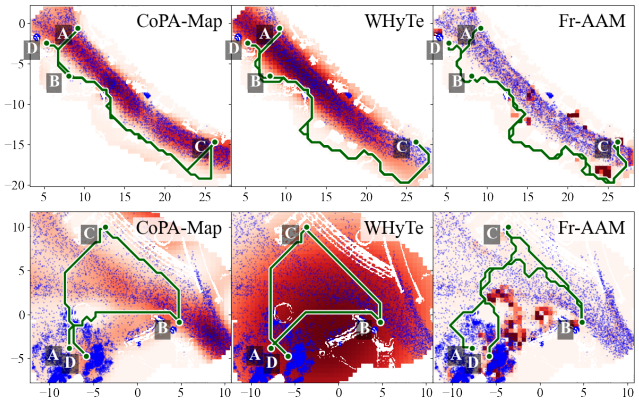


Fig. 5: Exemplary model predictions (intensity of red color scaled to respective maximum model output) and resulting paths (green) from the service disturbance experiment. The upper images represent the *hallway*, the lower images represent the *shops* scenario. Pedestrian data is shown as blue dots. Obstacles and areas outside the FOV are masked in white. Models requiring a grid representation were trained with resolution 0.5 m × 60 min.

ratio of $r = 1$ CoPA-Map leads to ca. 32 % less encounters over all paths compared to the Occupancy Map model for both the *hallway* and *shops* scenarios.

VI. Conclusion

We present CoPA-Map, a non-parametric method for spatio-temporal continuous modeling of human activity. Compared to other methods, CoPA-Map has advantages with respect to the quality of predictions and path planning, especially when pedestrian data is collected by moving robots. The model provides the basis for an extendable framework, that could also e.g. incorporate temporal trends through non-stationary covariance functions. As CoPA-Map is a single-output model, extensions need to be investigated to also incorporate movement direction and speed of pedestrians, e.g. by multi-output Gaussian Processes.

References

[1] T. Krajník, J. P. Fentanes, J. M. Santos, and T. Duckett, "FreMEn: Frequency Map Enhancement for Long-Term Mobile Robot Autonomy in Changing Environments," *IEEE Trans. Robot.*, vol. 33, no. 4, pp. 964–977, 2017.

[2] S. T. O'Callaghan, S. P. N. Singh, A. Alempijevic, and F. T. Ramos, "Learning navigational maps by observing human motion patterns," in *IEEE ICRA*, 2011, pp. 4333–4340.

[3] T. Vintr, Z. Yan, K. Eyisoy, F. Kubiš, J. Blaha, J. Ulrich, C. S. Swaminathan, S. Molina, T. P. Kucner, M. Magnusson *et al.*, "Natural criteria for comparison of pedestrian flow forecasting models," in *IEEE IROS*, 2020, pp. 11 197–11 204.

[4] M. Stuede, T. Lerche, M. A. Petersen, and S. Spindeldreier, "Behavior-Tree-Based Person Search for Symbiotic Autonomous Mobile Robot Tasks," in *IEEE ICRA*, 2021, pp. 2414–2420.

[5] T. Vintr, Z. Yan, T. Duckett, and T. Krajnik, "Spatio-temporal representation for long-term anticipation of human presence in service robotics," in *IEEE ICRA*, 2019, pp. 2620–2626.

[6] T. Vintr, S. Molina, R. Senanayake, G. Broughton, Z. Yan, J. Ulrich, T. P. Kucner, C. S. Swaminathan, F. Majer, M. Stachová *et al.*, "Time-varying pedestrian flow models for service robots," in *ECMR*. IEEE, 2019, pp. 1–7.

[7] M. Stuede, "CoPA-Map source code." [Online]. Available: https://github.com/MarvinStuede/copa-map.git

[8] R. Senanayake and F. Ramos, "Directional Grid Maps: Modeling Multimodal Angular Uncertainty in Dynamic Environments," in *IEEE IROS*, 2018, pp. 3241–3248.

[9] 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 RA-L*, vol. 2, no. 2, pp. 1093–1100, 2017.

[10] D. Ellis, E. Sommerlade, and I. Reid, "Modelling pedestrian trajectory patterns with Gaussian processes," *IEEE ICCV Workshops*, pp. 1229–1234, 2009.

[11] S.-Y. Chung and H.-P. Huang, "A Mobile Robot that Understands Pedestrian Spatial Behaviors," in *IEEE IROS*, 2010, pp. 5861–5866.

[12] S. T. O'Callaghan and F. T. Ramos, "Gaussian process occupancy maps," *Int. J. Robot. Res.*, vol. 31, no. 1, pp. 42–62, 2012.

[13] R. Senanayake, S. O'Callaghan, and F. Ramos, "Learning highly dynamic environments with stochastic variational inference," *IEEE ICRA*, pp. 2532–2539, 2017.

[14] R. Senanayake and F. Ramos, "Bayesian hilbert maps for dynamic continuous occupancy mapping," in *Conference on Robot Learning*, 2017, pp. 458–471.

[15] A. Tompkins and F. Ramos, "Fourier Feature Approximations for Periodic Kernels in Time-Series Modelling," in *AAAI Conference on Artificial Intelligence*, 2018.

[16] F. Jovan, J. Wyatt, N. Hawes, and T. Krajník, "A Poisson-Spectral Model for Modelling Temporal Patterns in Human Data Observed by a Robot," in *IEEE IROS*, 2016, pp. 4013–4018.

[17] S. Molina, G. Cielniak, and T. Duckett, "Go with the Flow: Exploration and Mapping of Pedestrian Flow Patterns from Partial Observations," *IEEE ICRA*, pp. 9725–9731, 2019.

[18] M. Lázaro-Gredilla and M. K. Titsias, "Variational heteroscedastic Gaussian process regression," in *ICML*, 2011.

[19] A. D. Saul, J. Hensman, A. Vehtari, and N. D. Lawrence, "Chained Gaussian Processes," *Proceedings of Machine Learning Research*, vol. 51, pp. 1431–1440, 2016.

[20] S. Kim and J. Kim, "Continuous occupancy maps using overlapping local gaussian processes," in *IEEE IROS*, 2013, pp. 4709–4714.

[21] D. J. MacKay *et al.*, "Introduction to Gaussian processes," *NATO ASI series F Computer and Systems Sciences*, vol. 168, pp. 133–166, 1998.

[22] A. Dutt and V. Rokhlin, "Fast Fourier transforms for nonequispaced data," *SIAM Journal on Scientific computing*, vol. 14, no. 6, pp. 1368–1393, 1993.

[23] H. Salimbeni, S. Eleftheriadis, and J. Hensman, "Natural gradients in practice: Non-conjugate variational inference in Gaussian process models," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2018, pp. 689–697.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] A. G. de G. Matthews *et al.*, "GPflow: A Gaussian Process Library using TensorFlow," *J. Mach. Learn. Res.*, vol. 18, no. 40, pp. 1–6, 2017.

[26] D. Brščić, T. Kanda, T. Ikeda, and T. Miyashita, "Person tracking in large public spaces using 3-D range sensors," *IEEE Trans. Hum. Mach. Syst.*, vol. 43, no. 6, pp. 522–534, 2013.

[27] S. Molina, G. Cielniak, and T. Duckett, "Robotic exploration for learning human motion patterns," *IEEE Trans. Robot.*, 2021.