Natural Criteria for Comparison of Pedestrian Flow Forecasting Models

Tomáš Vintr[1], Zhi Yan[5], Kerem Eyisoy[4], Filip Kubiš[1], Jan Blaha[1], Jiří Ulrich[1], Chittaranjan S. Swaminathan[2], Sergi Molina[3], Tomasz P. Kucner[2], Martin Magnusson[2], Gregorz Cielniak[3], Jan Faigl[1], Tom Duckett[3], Achim J. Lilienthal[2], Tomáš Krajník[1]

***Abstract*— Models of human behaviour, such as pedestrian flows, are beneficial for safe and efficient operation of mobile robots. We present a new methodology for benchmarking of pedestrian flow models based on the afforded safety of robot navigation in human-populated environments. While previous evaluations of pedestrian flow models focused on their predictive capabilities, we assess their ability to support safe path planning and scheduling. Using real-world datasets gathered continuously over several weeks, we benchmark state-of-the-art pedestrian flow models, including both time-averaged and time-sensitive models. In the evaluation, we use the learned models to plan robot trajectories and then observe the number of times when the robot gets too close to humans, using a predefined social distance threshold. The experiments show that while traditional evaluation criteria based on model fidelity differ only marginally, the introduced criteria vary significantly depending on the model used, providing a natural interpretation of the expected safety of the system. For the time-averaged flow models, the number of encounters increases linearly with the percentage operating time of the robot, as might be reasonably expected. By contrast, for the time-sensitive models, the number of encounters grows sublinearly with the percentage operating time, by planning to avoid congested areas and times.**

I. Introduction

In recent years, the maturity of robot localisation, mapping, obstacle avoidance, and motion planning methods enabled the creation of intelligent systems capable of reliable operation in structured environments. Some of these systems, like self-driving cars or delivery drones, already appeared on the market, while others were demonstrated in projects like STRANDS [1] or SPENCER [2]. Despite the readiness of the core methods above, there are few mobile robots operating routinely in human-populated environments. One of the possible causes is that, while robots are capable of reliable operation, they have trouble to be accepted by humans in the long-term. This issue is noticeable, especially after the initial excitement of working with robots ceases. The problem of long-term mobile robot acceptance was investigated in [3], where the authors conclude that a crucial factor is the way the robots navigate among humans. Robot behaviour using traditional navigation approaches is often considered inappropriate or aggressive [3].

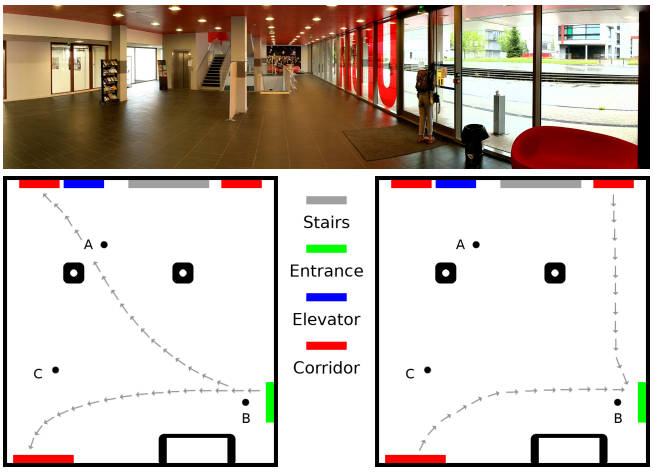

Fig. 1. Location where the experiments were performed. Photo and most prominent pedestrian flows during the morning (left) and evening (right).

Traditional path planning methods construct the robot path by considering the environment structure only and deal with unexpected obstacles in a reactive way. Typically, whenever a robot encounters an object not present in the original map, it alters the plan to avoid a collision. When moving around walking people, a robot is supposed to estimate their velocities and assess them when replanning its path [4]. However, the prediction accuracy of state-of-the-art methods does not allow the robot to plan accurate paths around humans in temporal horizons exceeding 1 second [5]. Moreover, the presence of a robot moving in a potentially colliding course causes people to alter their movement as well as forcing the robot to replan again. Such situations may result in confusing behaviour that people perceive negatively.

Instead of relying on reactive approaches, a robot can plan paths and schedule navigation by considering the typical movements of people learned from observations. In this way, a robot can schedule movements at times when locations are not crowded, and plan paths that either avoid or conform

[1] Artificial Intelligence Center, Czech Technical University. {name.surname}@fel.cvut.cz
[2] AASS Mobile Robotics and Olfaction Lab, Orebro University, Sweden {name.surname}@oru.se
[3] Lincoln Centre for Autonomous Systems (L-CAS), University of Lincoln, UK. {smolinamellado,gcielniak,tduckett}@lincoln.ac.uk
[4] Department of Computer Engineering, Faculty of Engineering, Marmara University, Turkey {name.surname}@gmail.com
[5] Distributed Artificial Intelligence and Knowledge Laboratory (CIAD), University of Technology of Belfort-Montbeliard (UTBM), France. zhi.yan@utbm.fr

The work has been funded by the OP VVV funded project CZ.02.101/0.0/0.0/16_019/0000765 “Research Center for Informatics”, CSF projects GA18-18858S and GC20-27034J, SGS19/176/OHK3/3T/13, FR-8J18FR018, PHC Barrande programme under grant agreement No. 40682ZH (3L4AV), European Union’s Horizon 2020 research and innovation programme under grant agreement No. 732737 (ILIAD), Toyota Partner Robot joint research project (MACPOLO).

with the previously observed flows of people. In order to achieve that, a robot needs to build environment models capable of capturing and forecasting the densities, directions, and velocities of pedestrian flow in its operational area. These models can be used to construct trajectories, which tend to be less intrusive to people, as they avoid congestion and other adverse situations [6]. The building, updating, and refining of models, which represent the spatial structure of pedestrian flows, was addressed in [6]–[10]. However, one can go beyond modelling of the spatial flow structure. As pointed out in [11], densities, velocities and directions of pedestrian flows vary over time. To address that, several recent contributions [12]–[16] have designed methods that can learn these variations from pedestrian data gathered by mobile robots.

Recently, authors of the works above compared their models in a joint paper [17] using two different statistics related to their predictive capabilities. While the paper [17] provides a fundamental insight into the models' capabilities, the authors conclude that another method of comparison is needed. Analogously to research in simultaneous localisation and mapping (SLAM), where method evaluation is not based primarily on the quality of the generated map, but rather on the accuracy of robot localisation using the map, we propose to evaluate these "flow maps" through the way they are actually used. Since the principal purpose in building pedestrian flow models is to enable unobstructive and safe path planning, we evaluate the models in terms of the *robot acceptance cost* of the generated paths, see Sec. III. We believe that robot acceptance can be approximated by counting the number of human-robot encounters, as these would force the robot and human to alter their path to avoid each other. Thus, in our experiments, we first train flow models on real-world data gathered for several weeks and let these models forecast a *robot acceptance cost map* of pedestrian movement for days not present in the training data, see Sec. III. Then, we use standard methods to generate robot trajectories that visit several locations in the environment and check if these trajectories intersect with the trajectories of pedestrians in the testing dataset. The number of encounters is then used to measure the utility of a spatio-temporal model for path planning.

We demonstrate, using the real-world datasets, that the proposed measures differentiate spatio-temporal models better than the usually used ones for this purpose. In the real robot experiment, we show that the proposed measures correlate with the experimental results, which indicates the interpretability of the quality scores. We also show that using human flow forecasts as priors for navigation tasks leads to more unobstructive navigation.

II. Related work

A fundamental capability of any autonomous mobile robot is navigation. The optimality of the planned path is strongly influenced by the quality of the robot's internal model of the surrounding environment. A substantial factor of the environment is its dynamics, which is related to the motion of people throughout the environment [7], [10], [12], [14]. The movement patterns, referred to as pedestrian or traffic flows, are influenced by the environment structure [18], time of the day [16], culture and other factors.

The work in [10] demonstrates that knowledge of these flows enables robots to move in a socially compliant manner, which has a positive impact not only on their navigation efficiency but also on their acceptance [3]. One of the first attempts to characterise the flows extended an occupancy grid model by gradual propagation of occupancy transitions from the adjacent cells [8]. A similar grid-based approach associated an input-output Markov model with each cell efficiently representing spatial relations of dynamics between them [7]. In [12], the authors discretise the movement of people in eight directions and associate each direction with a temporal model [19] capturing the periodic properties of the flows.

Other authors tried to model the pedestrian flows by continuous, rather than discrete representations. The paper [20] learns typical motions of people from long-term observations. The spatial layout of the flows is represented by Gaussian processes, and the authors show how their model can be used to plan robot motion. However, the approach did not consider the natural multi-modality of pedestrian flow distributions, e.g. bi-directional flows over a zebra crossing. A subsequent paper of the authors [21] elaborated on this particular aspect and presented an improved method capable of representing multi-modal distributions of pedestrian movement directions. To overcome the limitations of [8], the authors have proposed a continuous representation in [9], [22], which allows learning multi-modal models of flow directions and velocities from sparse data, using a set of semi-wrapped Gaussian mixture models. They demonstrated that their spatio-temporal pedestrian flow model improves the efficiency of motion planning for a nonholonomic robot in human crowds [6], [23].

However, the flows do not possess only spatial, but also temporal structure, because the variations of the flow densities, velocities and directions are subject to human habits. Thus, as shown in [12]–[14], [17], pedestrian flows can be modelled as cycloperiodic processes, i.e. their statistical properties vary with daily, weekly or other periods. In particular, the papers [12], [13] apply spectral-temporal models [19] to the individual cells of a discrete model. To overcome the problems of discretisation, the authors of [14], [17] employ the warped-hypertime paradigm [16] in their approaches. This paradigm represents the time not as a linear variable, but as a set of multiple dimensions wrapped into themselves. The aforementioned representation not only efficiently captures periodic variations of the pedestrian flow properties but also respects the spatio-temporal continuity of the flows.

A comparison of the aforementioned spatio-temporal pedestrian flow models was recently published in [17]. The predictive capabilities of the models were evaluated by two different criteria, root-mean-square error (RMSE) [24] and Chi-square distance [25]. Both methods create a multidimen-

sional histogram of measured values and compare them with the values predicted by a particular model. As demonstrated in [26], the resulting model rating may change with different grid resolutions when using these metrics. The results in [17] were inconclusive, as the values of RMSE and χ^2 distance did not vary significantly between the methods, and low RMSE and χ^2 distances were achieved by different methods. Inspired by SLAM evaluation approaches, which benchmark the quality of the maps by their ability to support accurate localisation, we propose to evaluate the flow models by their ability to support unobstructive navigation. Since the flow models aim to support the planning of safe paths, we propose to measure the quality of a given model by the number of encounters with humans, detected in the testing datasets. Similarly to [27], to reflect the ability of the models to represent the flow variations over time, we let our robots not only plan paths but also decide when is the best time to execute them if they have the opportunity to do so.

III. Acceptance-based Evaluation

The proposed evaluation method aims to quantify the model impact on the robot's navigation ability for social acceptance of a robot, as in [3]. The evaluation score is based on the traditional navigation paradigm used in mobile robotics, where the robot path is calculated by minimising some criterion, which is usually referred to as a cost. In order to optimise the path cost, path planning algorithms have to be able to retrieve the motion cost across each area of the robot's operational environment. The projection from a given position in space to a cost is referred to as a cost map. A typical cost map, which is based on the environment structure, encompasses the distance to obstacles, speed of motion, probability of successful traversal, and other aspects.

In our approach, we are trying to model the *robot acceptance cost* (RA cost) of robot navigation, which quantifies the discomfort level caused by the robot to the surrounding humans. For this purpose, we construct the cost map from the predictions of human flow models. The costs are derived from the likelihood of human flow at specific positions. These costs can be interpreted as the likelihood of forcing humans to alter their paths to avoid the robot presence at these specific positions. The cost map constructed in this way is denoted as a *robot acceptance cost map* (RA cost map).This RA cost map is then searched by traditional path planning methods to identify paths with a low probability of disrupting peoples' trajectories. Since a good pedestrian flow model should be able to generate a faithful RA cost map, it should significantly lower the number of encounters between a robot navigating through human-populated areas and humans moving towards their goals. Moreover, a model that captures how the flow properties change over time should be able to indicate when to traverse a given environment without disturbing too many people.

Based on the aforementioned ideas, we suggest two novel criteria for model quality evaluation. Both of them are calculated from a set of p imaginary robot navigation scenarios, each starting at a different time t_i, $i = 1 \dots p$, where a robot has a starting point and two different points that it needs to visit in order of its choice. To calculate the path, we create a cost map, where the costs are predicted by the pedestrian flow model as specified in Sec. IV-B. Then, we construct the path using Dijkstra's method, store its RA cost c_i, and transform the path into a trajectory setting the robot's speed to a constant. Then, using the detected human time-space positions retrieved from the test dataset, we calculate the number of *blind* robot-human encounters e_i during the robot movement.

Note that the aim of the evaluation methods is not to model human or robot behaviour during the encounter, but to create a tool to measure the difference between a model and the manifestation of the unknown underlying process. As such, robot and people are considered blind, entering the encounters with no reaction to it. The number of these blind robot-human encounters then expresses how many times the people or the robot would have to replan their movement.

A. Total Encounters

To obtain the first evaluation criterion, we count all blind robot-human encounters, referred to as *total encounters* TE during all planned passages through the environment,

$$TE = \sum_{i=1}^{p} e_i. \tag{1}$$

The value TE reflects the model's overall ability to support unobstructive planning by preventing the disruption of pedestrian flows. Lower total encounters mean better human flow model of the environment.

B. Expected Encounters

The second criterion reflects the similarity between a model and the spatio-temporal dynamics of the environment. We assume that a robot using a pedestrian flow model, which accurately represents how the flow intensities, directions and velocities change over time, should be able to schedule the service to navigate at times with a low number of encounters.

We again provide the robot with a set of navigation scenarios starting at t_i. We do not require that the robot performs every navigation task, but only a certain fraction of them – we refer to it as the *servicing ratio* $r \in [0,1]$. With a lower servicing ratio, the robot has more freedom to decide when to navigate through the area and when not. This reflects the situation when the robot has to perform a certain number of tasks during the day, but it can choose the best times to do so [27], [28].

Let us have a model that can predict RA path costs c_i for each planned path through the environment. We reorder these plans in ascending order by c_i and reindex them such that $k = 1 \dots p$ and $c_k \leq c_{k+1}$ for all k. Then we define *service disturbance* $E(s)$ as a sum of blind robot-human encounters e_k during the specific number s of planned passages as:

$$E(s) = \sum_{k=1}^{s} e_k. \tag{2}$$

Service disturbance corresponds to the situation when the robot chooses to traverse the environment only s times when it expects a low number of people, i.e. to avoid expected high RA path costs. Service disturbance for the methods investigated in this paper is shown in Fig. 2.

To characterise the ability of the method to predict the spatio-temporal dynamics of the environment, we define a function

$$Q(r) = E\left(\lfloor pr \rfloor\right), \tag{3}$$

which expresses the dependence of the service disturbance on the servicing ratio. Function $Q(r)$ is understood as a quantile function of the distribution of the service disturbance. (Note that total encounters $TE = Q(1) = E(p)$.) This distribution reflects the ability of the model to predict the fluctuation of the encounters in time. The second criterion of model quality is then defined as the expected value of the service disturbance:

$$EE = \int_0^1 Q(r)\, dr, \tag{4}$$

and we refer to this criterion as *expected encounters EE*.

The main reason for introducing both criteria, TE and EE, lies in a difference between time-averaged and time-sensitive models. For time-averaged models, EE has no meaning, and its value lies around $\frac{1}{2}TE$ (biased by the random ordering of services). Although there is apparent reason to compare time-sensitive models by TE, their sensitivity to the time-dependent changes should be compared by their ability to correctly predict the (relative) number of people at specific times, which leads to EE.

C. Details on path planning

For the sake of simplicity, the path planning problem is cast as a graph search over a two-dimensional grid with a cell dimension of 0.5 m. Nodes in the graph correspond to cells on the grid. A directed edge lies between every node in the graph and each of its eight neighbours on the grid. The costs of the directed edges are computed by incorporating predictions from the models of human flows. We use Dijkstra's algorithm to find a path from the start node to the goal that is least expensive in terms of the RA cost.

The models of human flows return the predicted number of people walking in some direction at some point in space-time or the likelihood of people going that direction. Construction of a cost map requires the models to calculate the most likely RA cost of a robot moving through a given cell in a particular direction.

Each scenario is inspired by a security robot procedure, where the robot has to visit a few predefined locations. In our case, there are one starting and finishing position (A) and two goal positions (B, C). The robot decides the order in which it will check the positions B and C, i.e. if it visits the locations in order (A, B, C, A) or (A, C, B, A). This decision is deduced from the mean of the costs belonging to each possible path – the robot chooses the path with the lower predicted cost.

To model people's personal space and the robot size, the path planning method assumes that the radius of the robot's social distance is 1 m, and encounter detection is triggered every 0.1 m along its trajectory. The robot speed was set to 0.5ms^{-1}. The speed, radius and cell size were chosen arbitrarily. The encounters are weighted similarly to the Extended Upstream Criterion [23]. This means that the highest value (2) represents the blind human-robot encounter when they are facing each other, while the lowest value (0) represents the encounter of human and robot moving in exactly the same direction. The robot performed 597 patrols during one day between 6:30 am and 8 pm starting every 80s. All models and code of the benchmark framework can be found online [29]. In the framework, it is possible to run tests with different speed, and radius of the robot, also there can be enabled non-uniform weights based on the directions of the flows during collisions.

D. Other Criteria

In addition to the proposed criteria to measure the quality of models, we included RMSE and Chi-square between the cost map predicted by the model and the ground truth obtained from the testing dataset, following the earlier comparison of human flow models [17]. RMSE is widely used in the time series forecasting:

$$RMSE = \sqrt{\frac{1}{p \cdot n \cdot a} \sum_{i=1}^{p} \sum_{q=1}^{n} \sum_{b=1}^{a} (x'_{i,q,b} - y'_{i,q,b})^2}, \tag{5}$$

normalised as a distribution over the testing space-time

$$\sum_{i=1}^{p} \sum_{q=1}^{n} \sum_{b=1}^{a} x'_{i,q,b} = \sum_{i=1}^{p} \sum_{q=1}^{n} \sum_{b=1}^{a} y'_{i,q,b} = 1, \tag{6}$$

and the Chi-square distance is used to compare histograms:

$$\chi^2 distance = \sum_{i=1}^{p} \sum_{q=1}^{n} \sum_{b=1}^{a} \frac{(x''_{i,q,b} - y''_{i,q,b})^2}{(x''_{i,q,b} + y''_{i,q,b})}, \tag{7}$$

normalised over the angles in every cell

$$\sum_{b=1}^{a} x''_{i,q,b} = \sum_{b=1}^{a} y''_{i,q,b} = 1 \tag{8}$$

where p is the number of scenarios, n is the number of positions, a is the number of angular bins, $x'_{i,q,b}$, resp. $x''_{i,q,b}$ is the normalised value of estimated cost for angle b at position q in scenario i, and $y'_{i,q,b}$, resp. $y''_{i,q,b}$ is the normalised value obtained from the ground truth at the identical position.

IV. Experimental evaluation

A. Evaluation dataset

The approaches described above were evaluated using a dataset collected in Building M at the University of Technology of Belfort-Montbeliard (UTBM). The data recording was performed by a Velodyne HDL-32E 3D lidar, using a reliable person detection method [30]. During the data collection, the lidar remained stationary in the reception room near the entrance of the building, which allowed to scan the main activity area of the hall, covering a total area of around 500 m^2 (Fig. 1). The data collection was performed on a

full 24/7 basis for three months in 2019, and the dataset also contains semantic information, including positions of the entrance, elevator, stairs, corridors etc.

For the purposes of this paper, we restricted the area of detections to cover only the main entrance hall (approx. 150 m^2). The training dataset includes eleven working days from March 2019 when the students were regularly attending classes, and the test dataset is the first Monday in April 2019. Each day contains approximately 300000 human detections. Every detection is represented by a vector $(t,x,y,\dot{x},\dot{y},\phi,v)$ – time of detection, $2d$ position, $2d$ velocity, orientation, and speed of the detected human. Note that $(\dot{x},\dot{y})$ and (ϕ,v) are mutually convertible.

The 3D lidar was mounted in a reception office at a height of about 1.2 m, providing a good overview of the environment for data collection (Fig. 1). The 3D point cloud generated by the Velodyne lidar was divided into separate sets using an adaptive clustering method [30]. These clusters were then classified as human or non-human using a support vector machine (SVM), and 2D positions of people were processed by a multi-target tracking method [31] based on a combination of Unscented Kalman Filter (UKF) and Nearest Neighbour Joint Probability Data Association method (NNJPDA). The trajectories calculated by the tracker were then examined manually, and outliers (e.g. static objects classified as people) were removed from the dataset.

B. Methods Involved in the Experiments

1) WHyTeS: The main idea of the WHyTeS is the projection of the data into the multidimensional vector space (*warped hypertime space*), where the data forms clusters. These clusters are understood as different time-dependent features of the measured phenomena [16]. The method combines density estimation and the spectral analysis tool FreMEn [19].

WHyTeS models the flows of people in space-time $(t,x,y,\dot{x},\dot{y})$, but for the testing method, we need to estimate the weights in the space of (t,x,y,ϕ). Therefore, we create a grid over $(t,x,y,\dot{x},\dot{y})$ and calculate for every ϕ the sum of predictions in cells that lie between $\phi-\frac{\pi}{8}$ and $\phi+\frac{\pi}{8}$ in every (t,x,y).

2) CLiFF-Map: Circular Linear Flow Field Map (CLiFF-Map) [9] is a technique for encoding motion patterns probabilistically. The probability density function (PDF) representing the CLiFF-Map is, for each point in the map, a semi-wrapped Gaussian mixture model (SWGMM).

The edge weights for the Dijkstra graph are computed using the Extended Upstream Criterion (EUC) proposed in [23]. The EUC penalises paths that do not conform to the underlying CLiFF-Map distribution. In the Dijkstra graph, the weight of each directed edge is computed using the CLiFF-Map components at the location of the node to which the edge leads.

3) STeF-Map: Spatio-Temporal Flow Map (STeF-Map) [13] is a representation that models the likelihood of motion directions on a grid-based map by a set of harmonic functions, which capture long-term changes of crowd movements over time. The underlying geometric space is represented by a grid, where each cell contains K temporal models, corresponding to K discretised orientations of people-motion through the given cell over time. The temporal models, which can capture patterns of people movement, are based on the FreMEn framework [19]. The edge weights for the Dijkstra graph are calculated using the EUC analogous to the CLiFF-Map application.

4) Other Models: In addition to the above listed state-of-the-art methods, we include for comparison three historical average based models and one time series forecasting model, all of which are commonly used as baseline methods. The average based models used were: *Means*, which predicts the mean of its past measurements for each spatial segment, and two histograms, *Histogram day* and *Histogram week*, that describe average day and week. Both histograms split their period into one-hour-long segments and compute the average value for each segment. The time series forecasting tool *Prophet* [32], was trained with measurements condensed into one-hour-long time steps. All these methods were trained on individual spatial segments without accounting for spatial relations. The last model, *Occupancy grid*, "predicts" at all cells a constant, which allows for the planning without forecasting the flow.

TABLE I
PERFORMANCE OF THE MODELS

Evaluated method	Total encounters	Expected encounters	χ^2 distance	Model error (RMSE)
Occupancy grid	7123	3993	8866	3.65e-06
Means	4185	2353	147769	4.08e-06
CLiFF-Map	3105	1480	144486	3.76e-06
Histogram week	5415	2857	80846	3.65e-06
Prophet	2637	596	59405	3.66e-06
WHyTeS	2835	385	82384	3.67e-06
STeF-Map	1548	332	70535	3.66e-06
Histogram day	2898	323	49981	4.02e-06

C. Expected Results of Models

We propose two criteria to quantify the quality of models, total encounters TE and expected encounters EE. The natural assumption is that time-averaged models (CLiFF-Map, Means) should have higher values of TE and EE compared to time-sensitive ones (WHyTeS, STeF-Map, Prophet, Histogram day, Histogram week) and the difference should be much more significant when comparing EE. Models using continuous time-forecasting (WHyTeS, STeF-Map, Prophet) should have a lower EE than discrete ones (Histogram day, Histogram week). Simple occupancy grid (planning without forecasting the flow) should achieve the significantly worst rating, while specialised CLiFF-Map should indicate better results than Means.

D. Evaluation results

The results of the evaluation, summarised in Table I, show that the proposed criteria, *total encounters* and *expected encounters*, can distinguish the quality of each model in a more

interpretative way compared to RMSE and χ^2distance. Total encounters are supposed to measure the similarity between a given model and the spatial dynamics of the environment, while expected encounters measures the similarity between a model and the spatio-temporal dynamics of the environment. We can see that the *Means* value of total encounters is significantly lower than the total encounters for *Occupancy grid*. The occupancy grid model "predicts" a constant value of the social cost over the whole grid, and the *Means* model predicts constant values for the social cost over the entire time. The specialised *CLiFF-Map* model is the best out of these three models which do not model the time. The expected encounters of these methods are very close to one-half of the total encounters, as expected. *CLiFF-Map* is better than *Means* in all of the criteria presented in Table I. However, the proposed criteria provide a more intuitive and interpretable comparison. Contrary to that, traditional comparison between *Occupancy grid* and the other two time-averaged models is counter-intuitive and misleading.

Except for the *Histogram week*, all models that predict the spatio-temporal dynamics of the environment reach significantly lower values of expected encounters than the time-averaged models. *Histogram day* achieved the best expected encounters, while *Histogram week* failed to learn usable model and its predictions from the perspective of the proposed criteria. To understand this result, we have to point out that the testing data comes from a working day and *Histogram day* benefits from that in the comparison. We can speculate with confidence that its result during the weekend would be weak because there are no lectures and human flows indicate very different behaviour. On the other hand, *Histogram week* can incorporate the difference between working days and weekends. However, a large number of temporal bins along with the short training period led to a poor model. Both *WHyTeS* and *STeF-Map* models covered three dominant periods in the data including a day and a week. *WHyTeS* covered also a period corresponding to the length of lectures during the working days, see Fig. 3. Sophisticated temporal modelling methods like *WHyTeS* and *STeF-Map* can compete with specialised histograms based on big data, while learning from significantly sparser training data sets [33]. The *Prophet* model showed that, for relatively short-term predictions, we can use regular time series forecasting methods and expect good results.

It is much harder, if not impossible, to interpret the characteristics of the predictors in the above way when analysing the results by RMSE or the χ^2 distance. This comparison supports our hypothesis that the proposed criteria characterise the predictive power of the spatio-temporal human flow models in a useful way. The popular and most common tools do not necessarily indicate how useful the analysed models are for unobstructive and safe path planning.

More in-depth insight into the dependence of the *service disturbance* on the *servicing ratio* (i.e. the dependence of encounters on the frequency of the traversals) is provided in Fig. 2. The graphs indicate that the *service disturbance* achieved by the time-averaged models, which neglect the

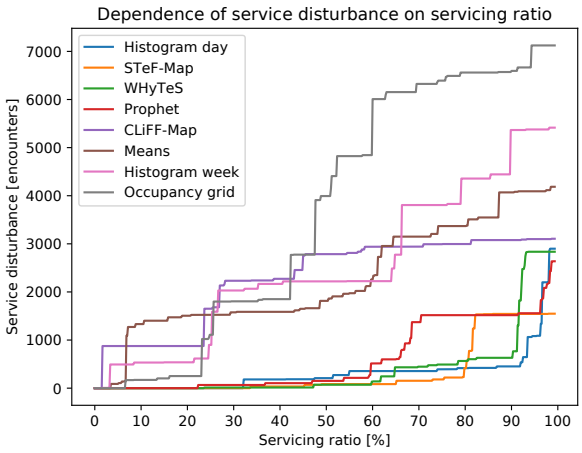

Fig. 2. The dependence of the number of encounters (service disturbance) on the frequency of the traversals (servicing ratio) achieved by different pedestrian flow models.

temporal variations of the flows, scales linearly with the *servicing ratio*. However, time-aware pedestrian flow models can identify times when the given area is crowded, and traversing through it will result in a large number of encounters. Thus, if allowed, these methods can choose not to enter the areas during these busy times.

In Fig. 2, the graphs indicate decreases of the service disturbance by about 80% for time-aware methods allowing the robot to drop 10-20% of tasks. Compared to traditional models, which capture the environment structure only, the well-trained flow representation can reduce the number of encounters by one order of magnitude at servicing ratio 90% and by order of two magnitudes at servicing ratio 50%.

The results also indicate that considering the spatio-temporal layout of the pedestrian flows has a tremendous impact on the number of people encountered, see total encounters of *STeF-Map* in Table I, when compared to the human flow uninformed baseline (Occupancy grid), which characterises only the static structure of the environment and associates each free cell with a fixed cost. Finally, we show that the *expected encounters*, which encompasses not only the ability to construct unobstructive and safe paths, but also the capability to support decisions when is the best time to execute them, is significantly lower for the models that explicitly represent the periodic variations of the pedestrian flow densities and directions.

E. Real-world experiment

To evaluate the impact of the proposed approach to unobstructive navigation, we performed two experiments during December 2019 at the entrance hall of the UTBM. The experiments were designed by researchers who do not work at UTBM, the robot used was never deployed at the experimental area before, and there was no advance notification to anyone involved in the experiment. The models used in the experiment were built from data collected during March 2019. Note that the models were trained on 14 days of data

TABLE II

COMPARISON OF REACTIVE AND ANTICIPATIVE NAVIGATION

Evaluated behaviour	Time	People total	People involved	People annoyed
Anticipative	9:20-10:00	115	17	0
Reactive	10:00-10:40	132	46	2
Anticipative	16:00-16:40	43	6	0
Reactive	16:40-17:20	23	14	1

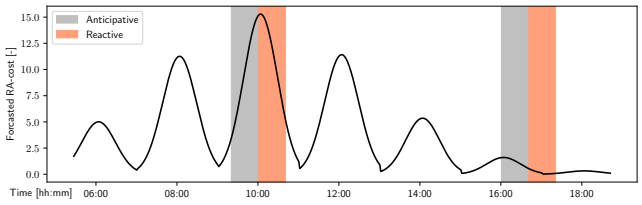

Fig. 3. Forecasted (8 month horizon) social cost and timeslots allocated for the experiment

collected 8 months in advance. Both experiments followed the methodology used in the previous section, Sec. III-C. In each experiment, we allocated two 40 minute slots to perform 10 patrols, during which the robot had to visit the 3 waypoints, see Fig. 1. The timeslots were chosen according to the forecasted RA path cost for a patrol performed at a particular time, see Fig. 3. The consecutive slots should have roughly the same number of people passing through the area.

The platform used was Toyota's HSR (Human Support Robot) robot [34], which was running Toyota's customised navigation method (undisclosable due to non-disclosure agreement) tailored for human-aware navigation. This method can detect and avoid people walking in the robot vicinity. The robot had to perform the patrols to minimise the time it took to visit the three waypoints.

During one of the timeslots, the robot used the occupancy grid map to plan its path. It had to perform the patrols uniformly distributed in time. Thus, the robot did not anticipate people presence, but it could avoid them reactively. During the other 40-minute timeslot, the robot scheduled its patrols so that the chance of interfering with people's trajectories would be minimised. It also chose the order of the waypoints, see Sec. III-C and Fig. 1, based on the predicted probabilistic model. In other words, the robot not only reacted to the people's presence, but it anticipated and adjusted its navigation plan and schedule accordingly in advance. The minimum time delay between two starts was set to two minutes, which corresponded to the estimated maximum time of travel.

While it is known that the perceived need to change one's course to avoid the robot is causing discomfort [3], we still decided to roughly assess the impact of the robot on the people passing through the area. To do so, we placed 3 paper sheets with removable tags near the 3 waypoints. These sheets asked the students to remove a tag if they felt that the robot was causing a nuisance by forcing them to avoid it. The idea behind this setting was to count how many people were so distracted by the robot that they perform an intentional operation as a reaction to the stressing situation.

Table II shows the results of the experiment. The number of people who walked through the hall during each time slot can be seen in the column 'People total'. The column 'People involved' indicates the number of people who walk through the area at the time as the robot. The values in the column 'People annoyed' corresponds to the number of tags removed during each time slot.

As the forecast, shown in Fig. 3, indicates, the number of people during both experiments was not uniformly distributed. There was about 70, 90, and 20 people walking through during 9:50–10:00, 10:00–10:15, and 16:00–16:10, respectively. The robot that scheduled its patrols, denoted as *Anticipative navigation*, chose to perform them in the intervals 9:30–9:50 and 16:20–16:40. Together with the values in the column 'People involved', it indicates that when using the spatio-temporal model as prior knowledge, the robot can effectively avoid the most congested times, which follows the results of the simulated experiments, Table I.

The evaluation of the acceptance of the robot by the involved pedestrians is far more tricky. We assumed that meeting fewer people and choosing the direction that suits the flows better lead to less annoyed people. The values in the column 'People annoyed' indicate that the hypotheses set in [3] were not disapproved.

V. CONCLUSION

We propose two new criteria, *Total encounters* and *Expected encounters* derived from the proposed *service disturbance* distribution, to measure the quality of the spatio-temporal models of human flows. Based on our previous experience from comparing flow models, we hypothesised that the traditional methods for their evaluation, based on measures of difference between the predicted and observed flows, e.g. root-mean-square error or χ^2 distance [13], [35], do not necessarily reflect the usefulness of these models for robots moving in human-populated environments. We argue that in the long-term, social acceptance of robots is negatively affected by events or "close encounters" where people feel the need to actively avoid a robot that moves in their direction [3]. The proposed criteria are derived from the estimated number of such events. The events are identified by planning the robot trajectory based on a particular flow model, and calculating their intersections with trajectories of humans in the testing datasets.

In the experiments, we show that contrary to the traditional methods, proposed criteria intuitively distinguish optimised human flow models. Moreover, a detailed analysis of *service disturbance* provides deep insight into the model's strong and weak aspects. We also show that models which accurately capture the spatio-temporal distributions of pedestrian flows allow for planning of trajectories which generate substantially less human-robot encounters than an uninformed flow baseline. On testing dataset, the time-sensitive models were able to reduce the number of encounters by order

of magnitude if the robot was allowed to drop 10% of the traversals during the busiest times and places, and by order of two magnitudes when dropping 60% of traversals. Our experiment with the real robot in the real environment indicates that scheduling and planning the tasks using the proposed *robot acceptance cost* map leads to more socially acceptable behaviour of a service robot.

In the future, we would like to investigate the balance between *robot acceptance* and other (e.g. time, distance) costs of navigation. Moreover, we want to investigate the influence of other system components, such as the path planner, on the *robot acceptance* cost of navigation. Furthermore, we would like to run more experiments with a real robot and assess the long-term acceptance of the robot by an ethnographic study in cooperation with psychologists.

REFERENCES

[1] N. Hawes *et al.*, "The strands project: Long-term autonomy in everyday environments," *IEEE Robotics & Automation Magazine*, vol. 24, no. 3, pp. 146–156, 2017.

[2] R. Triebel *et al.*, "Spencer: A socially aware service robot for passenger guidance and help in busy airports," in *Field and service robotics*. Springer, 2016, pp. 607–622.

[3] D. Hebesberger *et al.*, "A long-term autonomous robot at a care hospital: A mixed methods study on social acceptance and experiences of staff and older adults," *International Journal of Social Robotics*, vol. 9, no. 3, pp. 417–429, 2017.

[4] J. Müller *et al.*, "Socially inspired motion planning for mobile robots in populated environments," in *Proc. of International Conference on Cognitive Systems*, 2008.

[5] L. Sun *et al.*, "3dof pedestrian trajectory prediction learned from long-term autonomous mobile robot deployment data," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1–7.

[6] C. S. Swaminathan *et al.*, "Down the cliff: Flow-aware trajectory planning under motion pattern uncertainty," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 7403–7409.

[7] Z. Wang *et al.*, "Modeling motion patterns of dynamic objects by iohmm," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 1832–1838.

[8] T. P. Kucner *et al.*, "Conditional transition maps: Learning motion patterns in dynamic environments," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 1196–1201.

[9] ——, "Enabling flow awareness for mobile robots in partially observable environments," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 1093–1100, 2017.

[10] ——, *Probabilistic mapping of spatial motion patterns for mobile robots*. Springer, 2020.

[11] D. Brščić *et al.*, "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.

[12] S. Molina *et al.*, "Modelling and predicting rhythmic flow patterns in dynamic environments," in *Annual Conference Towards Autonomous Robotic Systems*. Springer, 2018, pp. 135–146.

[13] ——, "Go with the flow: Exploration and mapping of pedestrian flow patterns from partial observations," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 9725–9731.

[14] T. Vintr *et al.*, "Spatio-temporal representation for long-term anticipation of human presence in service robotics," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 2620–2626.

[15] W. Zhi *et al.*, "Spatiotemporal learning of directional uncertainty in urban environments with kernel recurrent mixture density networks," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4306–4313, 2019.

[16] T. Krajník *et al.*, "Warped hypertime representations for long-term autonomy of mobile robots," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3310–3317, 2019.

[17] T. Vintr *et al.*, "Time-varying pedestrian flow models for service robots," in *2019 European Conference on Mobile Robots (ECMR)*. IEEE, 2019, pp. 1–7.

[18] A. Rudenko *et al.*, "Human motion prediction under social grouping constraints," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 3358–3364.

[19] T. Krajník *et al.*, "Fremen: Frequency map enhancement for long-term mobile robot autonomy in changing environments," *IEEE Transactions on Robotics*, vol. 33, no. 4, pp. 964–977, 2017.

[20] S. T. O'Callaghan *et al.*, "Learning navigational maps by observing human motion patterns," in *2011 IEEE International Conference on Robotics and Automation*. IEEE, 2011, pp. 4333–4340.

[21] L. McCalman *et al.*, "Multi-modal estimation with kernel embeddings for learning motion models," in *2013 IEEE International Conference on Robotics and Automation*. IEEE, 2013, pp. 2845–2852.

[22] T. P. Kucner *et al.*, "Tell me about dynamics!: Mapping velocity fields from sparse samples with semi-wrapped gaussian mixture models," in *Robotics: Science and Systems Conference (RSS 2016), Workshop: Geometry and Beyond-Representations, Physics, and Scene Understanding for Robotics, University of Michigan, Ann Arbor, Ml, USA, June 18-22, 2016*, 2016.

[23] L. Palmieri *et al.*, "Kinodynamic motion planning on gaussian mixture fields," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 6176–6181.

[24] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International journal of forecasting*, vol. 22, no. 4, pp. 679–688, 2006.

[25] W. G. Cochran, "The $\chi 2$ test of goodness of fit," *The Annals of mathematical statistics*, pp. 315–345, 1952.

[26] F. Kubiš, "Application of spatiotemporal modeling used in robotics for demand forecast," B.S. thesis, Czech Technical University in Prague. Computing and Information Centre., 2020.

[27] J. P. Fentanes *et al.*, "Now or later? predicting and maximising success of navigation actions from long-term experience," in *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2015, pp. 1112–1117.

[28] L. Mudrova *et al.*, "An integrated control framework for long-term autonomy in mobile service robots," in *2015 European Conference on Mobile Robots (ECMR)*. IEEE, 2015, pp. 1–6.

[29] (2020) experiments for evaluating pedestrian flow forecasting models for socially compliant behaviour of mobile robots in populated environments. [Online]. Available: https://github.com/atomousek/evaluating_flow

[30] Z. Yan *et al.*, "Online learning for human classification in 3d lidar-based tracking," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 864–871.

[31] N. Bellotto and H. Hu, "Computationally efficient solutions for tracking people with a mobile robot: an experimental evaluation of bayesian filters," *Autonomous Robots*, vol. 28, no. 4, pp. 425–438, 2010.

[32] S. J. Taylor and B. Letham, "Forecasting at scale," *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018.

[33] J. Blaha, "Inferring temporal models of people presence from environment structrure," B.S. thesis, Czech technical university in Prague, 2020.

[34] T. Yamamoto *et al.*, "Development of human support robot as the research platform of a domestic mobile manipulator," *ROBOMECH journal*, vol. 6, no. 1, p. 4, 2019.

[35] R. Senanayake and F. Ramos, "Directional grid maps: modeling multimodal angular uncertainty in dynamic environments," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 3241–3248.