Temporal Smoothing for Joint Probabilistic People Detection in a Depth Sensor Network

Johannes Wetzel[1], Astrid Laubenheimer[1] and Michael Heizmann[2]

***Abstract*— Wide-area indoor people detection in a network of depth sensors is the basis for many applications, e.g. people counting or customer behavior analysis. Existing probabilistic methods use approximative stochastic inference to estimate the marginal probability distribution of people present in the scene for a single time step. In this work we investigate how the temporal context, given by a time series of multi-view depth observations, can be exploited to regularize a mean-field variational inference optimization process. We present a probabilistic grid based dynamic model and deduce the corresponding mean-field update regulations to effectively approximate the joint probability distribution of people present in the scene across space and time. Our experiments show that the proposed temporal regularization leads to a more robust estimation of the desired probability distribution and increases the detection performance.**

I. INTRODUCTION

Wide-area indoor people detection is a preprocessing task for a broad field of applications, such as people counting, customer behavior analysis, emergency detection in an ambient assisted living context or public security. Nonetheless, the vast majority of existing multi-view approaches use monocular video cameras and focus on pedestrian detection in outdoor scenarios, capturing the pedestrians from profile or frontal view. In this work we focus on the task of people detection in a network of low-cost commodity depth sensors. In contrast to the classical video surveillance scenario the mounting height is very limited in many indoor scenarios. This has three major implications: (i) the sensors capture the scene from the top-view to reduce occlusions in crowded scenes; (ii) since the resulting field of view of a single sensor is quite limited, sensor networks need to be employed to cover a larger area; (iii) position changes of people lead to drastically varying appearances as a result of the vertical top-view. In previous work [1] we propose a probabilistic framework which uses a generative scene model to leverage the full image evidence from all sensor views. For the final approximation of the probability distribution of people present in the scene a mean-field variational inference optimization is employed. However no temporal information is taken into account yet. A common way to leverage temporal information is to use the detections obtained from a isolated consecutive temporal frame as input for an off-the-shelf tracking-by-detection approach to get smooth person trajectories. Nonetheless, those methods do not take advantage of the full temporal information since the tracking component operates on a lossy representation of object detections and does not have access to the joint distribution of objects in the scene. In contrast, our goal is to avoid the loss of information by taking the image evidence from all sensor-views at every time step into account in order to approximate the joint probability distribution of people present in the scene across space and time.

In this work we investigate how the temporal context can be used to improve the detection performance by regularizing the underlying stochastic optimization process. Hence we present a novel extension of [1] which incorporates the temporal context given by a time series of multi-view depth observations (see Fig.1). The outcome of our approach could serve as input for tracking-by-detection post-processing to provide person trajectories. Our contribution is two-fold: (i) We present a probabilistic grid based dynamic model to define the joint distribution across space and time; (ii) we deduce the mean-field variational inference update equations to effectively approximate the desired probability distribution. In the evaluation we show that the proposed temporal regularization leads to a more robust approximation of the desired joint probability distribution and in consequence increases the detection performance.

II. RELATED WORK

The topic of indoor people detection in multiple overlapping depth images is not well studied yet. However, the related task of multi-view people detection and tracking with monocular video cameras has been studied in great detail [2], [3], [4]. In this section we will therefore discuss (i) methodically related work from the classical multi-view person tracking literature and (ii) existing approaches focusing on people detection and tracking in multi-view depth images. Many approaches in the literature address the problem of multi-view people detection and tracking by fusing local detections or tracklets into a global coordinate system. However, these approaches do not make full use of the multi-view image evidence, since the detection is performed on each single camera view independently. A related class of approaches uses generative modeling to jointly take advantage of the image evidence of all available views. For the detection of people Fleuret *et al.* [5] introduce the probabilistic occupancy

*The authors would like to thank the German Federal Ministry of Education and Research (BMBF), for funding the presented research under grant #13FH025IX6.

[1]Intelligent Systems Research Group (ISRG), Karlsruhe University of Applied Sciences, Karlsruhe, Germany johannes.wetzel@hs-karlsruhe.de
astrid.laubenheimer@hs-karlsruhe.de

[2]Institute of Industrial Information Technology (IIIT), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany michael.heizmann@kit.edu

 DOI: 10.1109/MFI49285.2020.9235267

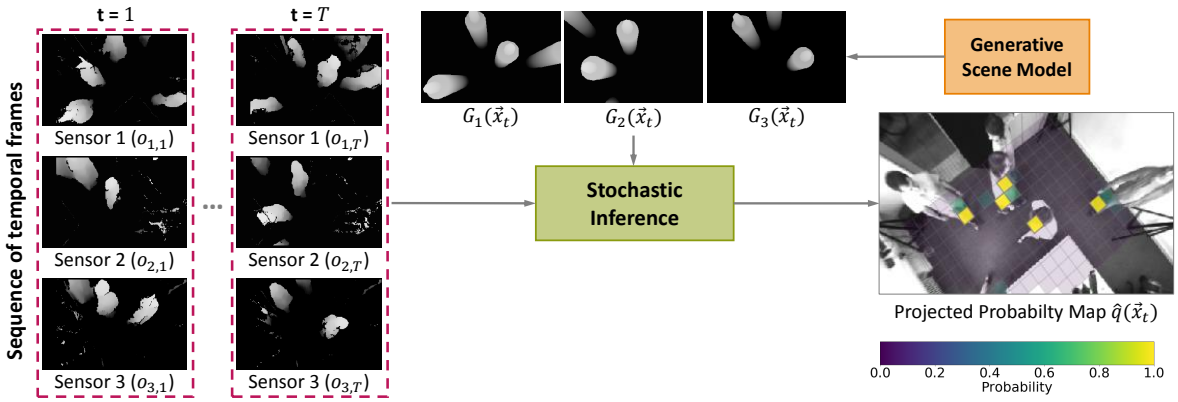


Fig. 1. Overview of the proposed approach. A time sequence of foreground segmented multi-view depth images from three sensors are used as input (left). The generative scene model generates synthetic depth images with respect to the given intrinsic and extrinsic sensor parameters (middle). The output of the stochastic inference are discrete probability maps representing the probability of people present on the ground floor across time and space (right).

map (POM). They use foreground-segmented binary images as input and employ a simple person model expressed by a rectangular bounding box to estimate probabilities of occupancy by mean-field variational inference. The POM is fed into a probabilistic tracking framework to provide smooth person trajectories. Alahi *et al.* [6] follow the idea of POM [5] and re-cast the task as a linear inverse optimization problem. Both methods utilize only a binary foreground mask as input and do not exploit the temporal context in the detection step. Baque *et al.* [7] introduce a state-of-the-art multi-view people detection architecture. They combine a generative scene model with a classical CNN architecture which additionally makes use of a Conditional Random Field (CRF) to resolve ambiguities arising from occlusion.

While the detection of people in a single depth image [8], [9], [10] has been intensively studied, only a few existing approaches address the problem of people detection in a network of depth sensors. Tseng *et al.* [11] present an indoor people detection and tracking system based on multiple active sensors in top-view. They fuse the point cloud of each sensor to a virtual global top-view depth image to get multi-view detections, which are fed into a tracking-by-detection scheme. Carraror *et al.* [12] propose an approach for human body pose estimation and tracking in a network of RGB-D sensors. In previous work [13] we re-cast the problem of people detection and tracking with multiple depth sensors as an inverse problem, employing an approximately differentiable scene model to detect people from arbitrary viewpoints. Following these ideas we introduced a probabilistic framework [1] based on a discrete scene configuration space. For stochastic inference a variational mean-field approximiation is used to jointly exploit the multi-view information in order to estimate the marginal probability distribution of people present in the scene.

The present work extends the framework proposed in [1] by taking the temporal context into account to approximate the full posterior distribution across space and time. While the mean-field variational inference method used in our approach is inspired by [5], we incorporate the temporal context jointly in the mean-field optimization to improve the detections across space and time.

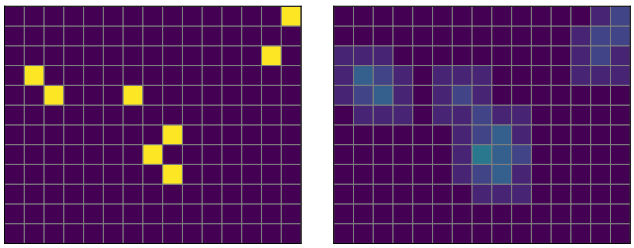
III. APPROACH

Our approach uses a time series of multi-view depth images as input and estimates the probability distribution of people present in a scene (see Fig.1). In previous work [1] we introduce a generative scene model to handle the different appearances of people due to the change of viewpoint as well as partial visibility of people e.g. due to occlusion or the limited field of view. The generative model is used in a probabilistic framework which leverages the full multi-view information given in the overlapping image regions for joint probabilistic people detection by using mean-field variational inference. In this work we extend the method introduced in [1] by taking the temporal context into account, thus defining the full joint distribution across all sensor views and time steps (Sect. III-A). Therefore, we introduce a dynamics model (Sect. III-A.2) to express the probability flow over time. We deduce how mean-field variational inference can be used to approximate the desired joint probability distribution effectively to estimate the marginal probabilities of people present in the scene across space and time (Sect. III-B).

A. Probabilistic Model

We assume that the intrinsic and extrinsic camera parameters and the common ground floor plane are known from the initial calibration. The ground floor area is discretized into a 2D-grid of n locations for each time step t with $1 \leq t \leq T$. Each location u_i will be assigned a realization $x_{i,t}$ of a Bernoulli random variable $X_{i,t} \sim \mathcal{B}(\mu)$, where μ denotes the probability of a person present at location u_i at time t. The scene configurations for one time step t are given as the vector $\mathbf{x}_t = (x_{1,t}, \ldots, x_{n,t})^T \in \{0,1\}^n$ (see Fig. 2(a)). The foreground-segmented depth observations at the corresponding time step are given as $\mathbf{o}_t = (o_{1,t}, \ldots, o_{c,t})^T$, acquired from depth sensors $S_1 \ldots S_C$. The joint probability distribution for time steps $1..T$ is given as

$$p(\mathbf{x}_{1:T}, \mathbf{o}_{1:T}) = p(\mathbf{o}_{1:T}|\mathbf{x}_{1:T})p(\mathbf{x}_{1:T}). \quad (1)$$



(a) $\mathbf{x}_{t-1}$ (b) $p(\mathbf{x}|\mathbf{x}_{t-1})$

Fig. 2. Example of proposed discrete dynamics model for one scene configuration (a) with corresponding output distribution (b) for $\mu_{self} = 0.2$ and $w_k = 1$.

Assuming conditional independence of the observations over time $p(\mathbf{o}_{1:T}|\mathbf{x}_{1:T}) = \prod_{t=1}^{T} p(\mathbf{o}_t|\mathbf{x}_t)$ and first order Markov dynamics $p(\mathbf{x}_t|\mathbf{x}_{1:t-1}) = p(\mathbf{x}_t|\mathbf{x}_{t-1})$ the posterior distribution can be written as

$$p(\mathbf{x}_{1:T}|\mathbf{o}_{1:T}) = \frac{\prod_{t=1}^{T} p(\mathbf{o}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{x}_{t-1})}{p(\mathbf{o}_{1:T})}, \tag{2}$$

with $p(\mathbf{x}_1|\mathbf{x}_o)p(\mathbf{x}_o) = p(\mathbf{x}_1)$.

1) Data Likelihood: Since we use the same definition for the likelihood $p(\mathbf{o}_t|\mathbf{x}_t)$ as introduced in [1] we only briefly recap the definition in this section. Assuming that the observations are conditionally independent given a scene configuration $\mathbf{x}_t$ the data likelihood can be written as

$$p(\mathbf{o}_t|\mathbf{x}_t) = \prod_{c=1}^{C} p(o_{c,t}|\mathbf{x}_t). \tag{3}$$

To handle the different appearance of people due to the change in viewpoint, a generative scene model $G_c(\mathbf{x}_t, \mathbf{P}_c)$ is employed, which maps a scene configuration $\mathbf{x}_t$ and a given projection matrix $\mathbf{P}_c$ to a synthetic observation (i.e. synthetic depth image) from the perspective of sensor S_c (see Fig. 1 top middle). We assume that our given observations suffer from Gaussian noise, yielding an observation likelihood

$$p(o_{c,t}|\mathbf{x}, \sigma) \propto \exp\left(-\frac{1}{2\sigma^2}\left\|o_{c,t} - G_c(\mathbf{x}, \mathbf{P}_c)\right\|_2^2\right). \tag{4}$$

2) Dynamics Model: Since we do not focus on tracking but on leveraging the temporal context to regularize the stochastic optimization (see Sect. III-B), we propose a rather simple, grid based dynamics model without modeling explicit motion of objects. This leads to a computationally feasible model which represents the flow of probability across time and space. For the sake of simplicity we assume that the probability of a scene configuration $\mathbf{x}_t$ given the previous state $\mathbf{x}_{t-1}$ factorizes as

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) = \prod_{j=1}^{n} p(x_{j,t}|\mathbf{x}_{t-1}). \tag{5}$$

To express the distribution $p(x_{j,t}|\mathbf{x}_{t-1})$ as a weighted sum of all previous $x_{i,t-1}$ being in state one (meaning that a person is present), we introduce a random variable $\mathbf{Z}$ with the realizations being $\mathbf{z} = (z_1, \ldots, z_n)^T$ with $z_k \in \{0, 1\}$ and one-hot-encoding such that $\sum_{k=1}^{n} z_k = 1$. Introducing $\mathbf{z}$ to the distribution $p(x_{j,t}|\mathbf{x}_{t-1})$ leads to the joint distribution

$$p(x_{j,t}, \mathbf{z}|\mathbf{x}_{t-1}) = p(x_{j,t}|\mathbf{z}, \mathbf{x}_{t-1})p(\mathbf{z}|\mathbf{x}_{t-1}). \tag{6}$$

Since $\mathbf{z}$ is one-hot-encoded it follows

$$\begin{aligned} p(x_{j,t}|\mathbf{z}, \mathbf{x}_{t-1}) &= \prod_{k=1}^{n} p(x_{j,t}|z_k = 1, \mathbf{x}_{t-1})^{z_k} \\ &= p(x_{j,t}|z_k = 1), \end{aligned} \tag{7}$$

where $p(x_{j,t}|z_k = 1)$ reflects the probability of $x_{j,t}$ given that one particular cell at index k in the previous state $\mathbf{x}_{t-1}$ is one. Marginalization over $\mathbf{z}$ gives the mixture model

$$\begin{aligned} p(x_{j,t}|\mathbf{x}_{t-1}) &= \sum_{\mathbf{z}\in\{0,1\}^n : |\mathbf{z}|=1} p(x_{j,t}, \mathbf{z}|\mathbf{x}_{t-1}) \\ &= \sum_{k=1}^{n} p(x_{j,t}|z_k = 1)p(z_k = 1|\mathbf{x}_{t-1}), \end{aligned} \tag{8}$$

where the distribution $p(z_k = 1|\mathbf{x}_{t-1})$ can be interpreted as normalization weights, such that

$$p(z_k = 1|\mathbf{x}_{t-1}) = \begin{cases} w_k = \frac{1}{||\mathbf{x}_{t-1}||_1}, & \text{if } x_{k,t-1} = 1 \\ 0, & \text{else.} \end{cases} \tag{9}$$

Finally the probability flow depends on the transition distribution $p(x_{j,t}|z_k = 1)$ which denotes the probability that a person is present at location u_j given that a person was present at the previous time step at location u_k. Since in our setup only little movement per time step is expected we assume that a person will stay on the current location with a specific probability μ_{self} and moves to one of its direct eight neighbor cells uniformly (see Fig. 2). Hence we define

$$p(x_{j,t}|z_k = 1) = \begin{cases} \mathcal{B}(x_{j,t}|\mu_{self}) \ , & \text{if } j = k \\ \mathcal{B}(x_{j,t}|\mu_{ne}) \ , & \text{if } j \in ne(k) \\ 1 - x_{j,t} \ , & \text{else,} \end{cases} \tag{10}$$

with $ne(k)$ being the set of the direct neighbors of u_k. We define $\mu_{ne} = (1 - \mu_{self})/|ne(k)|$, which leads to the special case where the emitted probability for one person present equals to one. Notice that because of the normalization weights w_k it is also possible to use more sophisticated transition probability distributions. Another consequence for this special choice of μ_{ne} is that we can set $w_k = 1$ while (8) still meets the requirements of a probability density function. This has the side effect that the expected number of people in the scene with respect to the dynamics model stays constant.

B. Temporal Regularization of Stochastic Inference

Since the posterior (2) is intractable due to dimensionality of the latent space we use mean-field variational inference [14] to approximate the distribution $p(\mathbf{x}_{1:T}|\mathbf{o}_{1:T})$ by a simpler proxy distribution $q(\mathbf{x}_{1:T})$. The optimization objective is given as

$$\hat{q}(\mathbf{x}_{1:T}) = \arg\min_q \mathrm{KL}(q(\mathbf{x}_{1:T}) \,||\, p(\mathbf{x}_{1:T}|\mathbf{o}_{1:T})). \tag{11}$$

We assume a fully-factorized proxy distribution $q(\mathbf{x}_{1:T}) = \prod_{i=1}^{n}\prod_{t=1}^{T} q_{i,t}(x_{i,t})$ where each $q_{i,t}(x_{i,t})$ denotes the

marginal probability of a person present at location u_i at time step t. Let $\langle\cdot\rangle_{h(x)}$ be the expected value with respect to a distribution $h(x)$ and $q(\mathbf{x}_{1:T} \setminus x_{i,t}) = \prod_{j=1:j\neq i}^{n}\prod_{k=1:k\neq t}^{T} q_{j,k}(x_{j,k})$ the mean-field distribution without the element $x_{i,t}$. According to the general mean-field equation (see [15, 625 ff.]) the optimal update with respect to the objective (11) is given as

$$q_{i,t}(x_{i,t}) \propto \exp\Big(\langle \log p(\mathbf{x}_{1:T}|\mathbf{o}_{1:T})\rangle_{q(\mathbf{x}_{1:T}\setminus x_{i,t})}\Big). \quad (12)$$

Considering that each $x_{i,t}$ is Bernoulli distributed, the final update for $x_{i,t}$ being in state one is given as

$$q_{i,t}(x_{i,t}=1) = \left[1+\exp\left(E_{i,t}\right)\right]^{-1}. \quad (13)$$

For more detailed derivation we refer to previous work [1, Sect. 3(B)]. Inserting the probabilistic model defined in Sect. III-A the update expectation in (13) expands to

$$\begin{aligned} E_{i,t} &= \left\langle \log \frac{p(\mathbf{o}_{1:T},\mathbf{x}_{1:T}|x_{i,t}=0)}{p(\mathbf{o}_{1:T},\mathbf{x}_{1:T}|x_{i,t}=1))} \right\rangle_{q(\mathbf{x}_{1:T}\setminus x_{i,t})} \\ &= \left\langle \log \frac{\prod_{k=1}^{T} p(\mathbf{o}_k|\mathbf{x}_k, x_{i,t}=0)}{\prod_{k=1}^{T} p(\mathbf{o}_k|\mathbf{x}_k, x_{i,t}=1)} \right. \\ &\quad \left. + \log \frac{\prod_{k=1}^{T} p(\mathbf{x}_k|\mathbf{x}_{k-1}, x_{i,t}=0)}{\prod_{k=1}^{T} p(\mathbf{x}_k|\mathbf{x}_{k-1}, x_{i,t}=1)} \right\rangle_{q(\mathbf{x}_{1:T}\setminus x_{i,t})}. \end{aligned} \quad (14)$$

Using the linearity of expectation we can express (14) as the sum of a data and a temporal expectation

$$E_{i,t} = E_{i,t}^{\text{data}} + \underbrace{E_{i,t}^{\text{past}} + E_{i,t}^{\text{future}}}_{\text{temporal exp.}}. \quad (15)$$

All terms in (14) which are independent of $x_{i,t}$ cancel out, therefore the data term can be isolated to

$$E_{i,t}^{\text{data}} = \left\langle \log \frac{p(\mathbf{o}_t|\mathbf{x}_t, x_{i,t}=0)}{p(\mathbf{o}_t|\mathbf{x}_t, x_{i,t}=1)} \right\rangle_{q(\mathbf{x}_t\setminus x_{i,t})}. \quad (16)$$

Since we focus on the temporal regularization in this work, we refer to previous work [1] for further elaboration on the efficient approximation of the data term. Inserting the dynamics model (5) in (14) and again using the fact that all terms which are independent of $x_{i,t}$ cancel out, it turns out that the temporal part of the expectation can be separated into a part

$$\begin{aligned} E_{i,t}^{\text{past}} &= \left\langle \log \frac{p(\mathbf{x}_t|\mathbf{x}_{t-1}, x_{i,t}=0)}{p(\mathbf{x}_t|\mathbf{x}_{t-1}, x_{i,t}=1)} \right\rangle_{q(\mathbf{x}_t\setminus x_{i,t},\mathbf{x}_{t-1})} \\ &= \left\langle \log \frac{\prod_j p(x_{j,t}|\mathbf{x}_{t-1}, x_{i,t}=0)}{\prod_j p(x_{j,t}|\mathbf{x}_{t-1}, x_{i,t}=1)} \right\rangle_{q(\mathbf{x}_t\setminus x_{i,t},\mathbf{x}_{t-1})} \\ &= \left\langle \log \frac{1-p(x_{i,t}=1|\mathbf{x}_{t-1})}{p(x_{i,t}=1|\mathbf{x}_{t-1})} \right\rangle_{q(\tilde{\mathbf{x}}_{i,t-1})} \end{aligned} \quad (17)$$

which does only depend on the previous state $\mathbf{x}_{t-1}$ and a slightly more evolved part

$$\begin{aligned} E_{i,t}^{\text{future}} &= \left\langle \log \frac{p(\mathbf{x}_{t+1}|\mathbf{x}_t, x_{i,t}=0)}{p(\mathbf{x}_{t+1}|\mathbf{x}_t, x_{i,t}=1)} \right\rangle_{q(\mathbf{x}_{t-1:t+1}\setminus x_{i,t})} \\ &= \left\langle \log \frac{\prod_j p(x_{j,t+1}|\mathbf{x}_t, x_{i,t}=0)}{\prod_j p(x_{j,t+1}|\mathbf{x}_t, x_{i,t}=1)} \right\rangle_{q(\mathbf{x}_t\setminus x_{i,t},\mathbf{x}_{t+1})} \\ &= \left\langle \sum_{j\in ne(i)} \log \frac{p(x_{j,t+1}|\mathbf{x}_t, x_{i,t}=0)}{p(x_{j,t+1}|\mathbf{x}_t, x_{i,t}=1)} \right\rangle_{q(\tilde{\mathbf{x}}_{i,t},\tilde{\mathbf{x}}_{i,t+1})} \end{aligned} \quad (18)$$

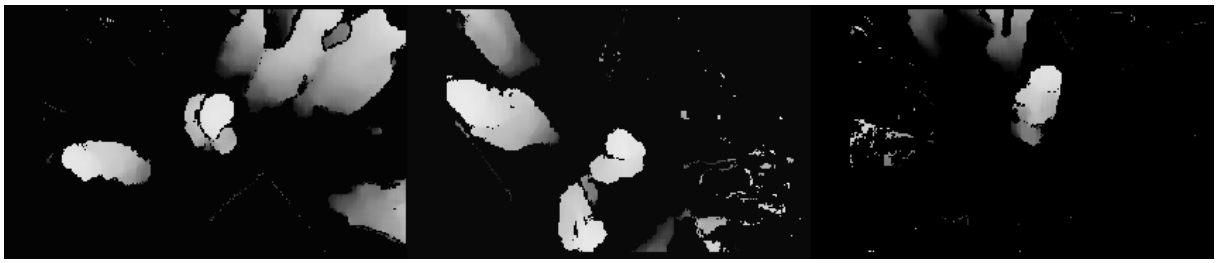
which depends on the current state $\mathbf{x}_t$ and the future state in the next time step $\mathbf{x}_{t+1}$. Since our dynamics model operates only on a local neighborhood (see (10)) we only need to consider the reduced neighborhood scene configurations $\tilde{\mathbf{x}}_{i,t} \in \{0,1\}^8$, making the estimation of the expectations (17,18) computationally feasible.

C. Implementation Details

For the mean-field optimization we use asynchronous coordinate-ascent variational inference (CAVI) [14], thus all $q_{i,t}(\cdot)$ are updated sequentially according to (13) with respect to the previous mean-field state $q(\mathbf{x}_{1:T} \setminus x_{i,t})$. In each iteration the time slices $q(\mathbf{x}_t)$ are consecutively updated from $1..T$. This implies that the temporal context does have a direct impact on the estimation of the data term on the next iteration since the mean-field distribution after one iteration is effected jointly by all temporal frames. Since the future term (18) relies on the mean-field state from the next time step we disable the future term in the first iteration. To take into account that people can enter the observable area we initialize all border grid cells with a probability of $q_{\text{init}}(x_i) = 0.5$ with $i \in$ `Border`. To weight the temporal terms we extend (15) to $E_{i,t} = E_{i,t}^{\text{data}} + \beta E_{i,t}^{\text{past}} + \gamma E_{i,t}^{\text{future}}$ with $\beta, \gamma \in [0,1]$. For the approximation of the data term (16) the asymmetric L1-image distance introduced in [1, Eq. (12)] is used.

IV. EVALUATION

We analyze the effects of using the temporal context as proposed in Sect. III on a sequence of 600 annotated consecutive temporal frames. Each temporal frame contains three foreground segmented depth images with a resolution of 376×240 pixel, recorded from three commodity stereo-vision-based depth sensors (see Fig. 3(a)). The foreground segmentation is obtained by simple static background subtraction. The sensors have a top view on the scene and are mounted at a height of three meters, having fields of view with significant overlap. For the evaluation of our approach we use a discrete ground floor grid with 15×12 grid points, corresponding to a horizontal and vertical distance of 33 cm between adjacent grid points. For the quantitative evaluation we use the precision-recall metric, where the precision is given by $TP/(TP+FP)$ and the recall by $TP/(TP+FN)$; TP, FP, FN are the counts of the true positives, false positives and false negatives, respectively. The F1-Score



(a) Input depth observations at one time step from three sensors (temporal frame)

(b) No temporal context used

(c) With temporal context

Fig. 3. Exemplary mean-field optimization results depicted for one temporal frame (a). (b,c) show the resulting marginal probability map projected onto the ground floor, false negatives are marked with a red dot.

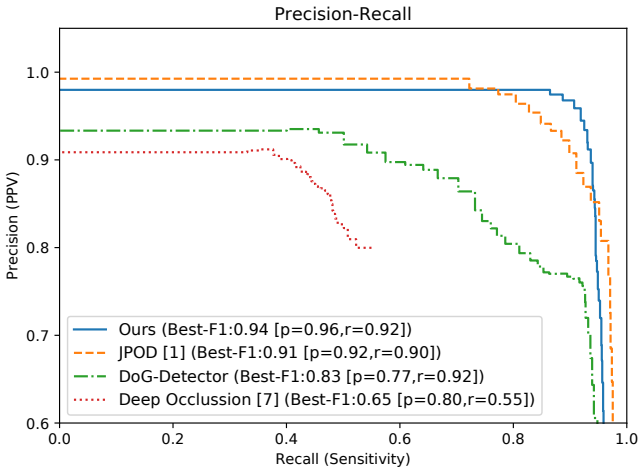
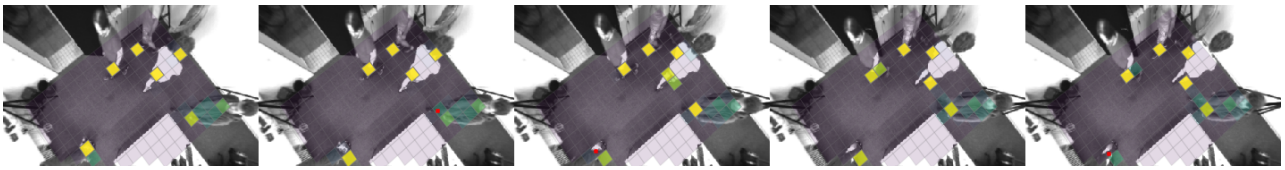


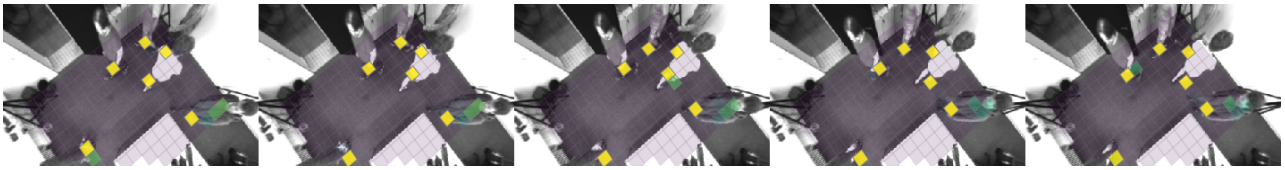
Fig. 4. Precision-Recall curves showing the performance of our approach with and without temporal context.

is defined as the harmonic mean of precision and recall, $F_1 = (2 \times \text{precision} \times \text{recall})/(\text{precision} + \text{recall})$. During evaluation we observed that the influence of the proposed future term in (18) on the quantitative results is negligible for the proposed update strategy and motion model. Therefore the presented results are based only on the past and data term defined in (16,17) respectively. For evaluation we run six mean-field iterations with the parameters $\beta = 0.65$, $\gamma = 0.0$, $\mu_{self} = 0.8$ and $w_k = 1$.

Fig. 4 shows the precision-recall performance for the proposed method compared to: (i) the joint probablistic people detection method (JPOD) without temporal context as introduced in [1]; (ii) a difference of Gaussian blob detector (DoG-Detector) which is applied on the foreground segmented depth images of each sensor independently and the finaly detections are obtained by proximity clustering on the ground plane; (iii) Deep Occlussion [7], a SOTA deep learning architecture for multi-view people detection (for evaluation we use the available pre-trained model; as input we stack the grey scale observations to a three channel image to be compatible with the RGB architecture). The top-view data set is challenging for Deep Occlusion since it was trained with RGB images containing people in the profile-view and is now applied to top-view grey scale images. Notice that the mentioned approaches (i-iii) operate on a single temporal frame and do not make use of the given temporal context. Although the performance of our previous work JOPD [1] without temporal context is already quite high (best F1-score of 0.91) the results show that the



(a) No temporal context



(b) With temporal context

Fig. 5. Mean-field results for five consecutive frames, projected into sensor view one.

exploitation of temporal context can increase the overall precision and recall performance (best F1-score of 0.94).

Fig. 3(b), (c) and Fig. 5 illustrate mean-field optimization results. The final marginal probability distribution $\hat{q}(\mathbf{x}_t)$ is projected onto the ground floor, where purple correspond to a probability of zero and yellow to one respectively (see Fig. 1 left bottom for the color scale). In Fig. 3 an exemplary temporal frame with a typical positive effect of the temporal regularization is shown. Without temporal context (Fig. 3(b)) the estimated marginal probability distribution contains high uncertainty around two targets due to partial visibility and measurement noise, leading to two false negatives. In contrast, exploiting the temporal context can resolve those uncertainties, leading to a marginal distribution with clean peaks (Fig. 3(c)). Similar effects can be observed in Fig. 5, where a short sequence of mean-field results (shown only in sensor view one) is depicted. On a single CPU core[1], our non-optimized Python implementation needs approximately $700\,ms$ per temporal frame. We observed that the run time per frame decreases slightly on average by using the temporal context. This can be explained by the fact that the proposed dynamics model effectively restricts the set of grid cells where a person can be present with a probability greater than zero, thus less mean-field updates need to be evaluated.

V. CONCLUSION

In this work we have presented a novel extension for probabilistic people detection in a depth sensor network, which leverages the temporal context to regularize the stochastic mean-field optimization process. We proposed a probabilistic grid based dynamics model and deduced the corresponding mean-field update equations to effectively approximate the joint distribution of people present in the scene across space and time. Our results have shown that the introduced temporal regularization leads to a more robust estimation of the desired joint probability distribution and in consequence increases the detection performance.

Future work will focus on extended quantitative evaluation as well as the investigation of more sophisticated grid based dynamics models and their influence on the temporal regularization.

[1]Intel Core-i7@2.9Ghz

REFERENCES

[1] J. Wetzel, A. Laubenheimer, and M. Heizmann, "Joint probabilistic people detection in overlapping depth images," *IEEE Access*, vol. 8, pp. 28349–28359, 2020.

[2] L. Hou, W. Wan, J. N. Hwang, R. Muhammad, M. Yang, and K. Han, "Human tracking over camera networks: a review," *EURASIP J. Adv. Signal Process.*, vol. 2017, p. 43, Dec. 2017.

[3] R. Iguernaissi, D. Merad, K. Aziz, and P. Drap, "People tracking in multi-camera systems: a review," *Multimed. Tools Appl.*, vol. 78, pp. 10773–10793, Apr. 2019.

[4] X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern Recognit. Lett.*, vol. 34, pp. 3–19, Jan. 2013.

[5] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 267–282, 2008.

[6] A. Alahi, L. Jacques, Y. Boursier, and P. Vandergheynst, "Sparsity driven people localization with a heterogeneous network of cameras," *J. Math. Imaging Vis.*, vol. 41, no. 1-2, pp. 39–58, 2011.

[7] P. Baque, F. Fleuret, and P. Fua, "Deep occlusion reasoning for multi-camera multi-target detection," in *Proc. IEEE Int. Conf. Comput. Vision, ICCV*, vol. 2017-Octob, pp. 271–279, 2017.

[8] L. Tian, M. Li, Y. Hao, J. Liu, G. Zhang, and Y. Q. Chen, "Robust 3-d human detection in complex environments with a depth camera," *IEEE Trans. Multimed.*, vol. 20, pp. 2249–2261, Sept. 2018.

[9] V. Carletti, L. Del Pizzo, G. Percannella, and M. Vento, "An efficient and effective method for people detection from top-view depth cameras," in *Proc. IEEE Int. Conf. Adv. Video Signal Based Surveillance, AVSS*, Oct. 2017.

[10] C. Ertler, H. Possegger, M. Opitz, and H. Bischof, "Pedestrian detection in rgb-d images from an elevated viewpoint," in *Proc. 22nd Comput. Vis. Winter Work.*, TU Wien, Pattern Recongition and Image Processing Group, 2017.

[11] T. E. Tseng, A. S. Liu, P. H. Hsiao, C. M. Huang, and L. C. Fu, "Real-time people detection and tracking for indoor surveillance using multiple top-view depth cameras," in *Proc. IEEE Int. Conf. Intell. Robot. Syst. IROS*, pp. 4077–4082, 2014.

[12] M. Carraro, M. Munaro, J. Burke, and E. Menegatti, "Real-time marker-less multi-person 3d pose estimation in rgb-depth camera networks," in *Adv. Intell. Syst. Comput.*, vol. 867, pp. 534–545, Springer, Cham, June 2019.

[13] J. Wetzel, S. Zeitvogel, A. Laubenheimer, and M. Heizmann, "Towards global people detection and tracking using multiple depth sensors," in *Proc. IEEE Int. Symp. Electron. Telecommun. ISETC*, pp. 1–4, Nov. 2018.

[14] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *J. Am. Stat. Assoc.*, vol. 112, no. 518, pp. 859–877, 2017.

[15] D. Barber, *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2011.