IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 16, 2020, accepted February 2, 2020, date of publication February 6, 2020, date of current version February 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2972055

Joint Probabilistic People Detection in Overlapping Depth Images

JOHANNES WETZEL[1], ASTRID LAUBENHEIMER[1], AND MICHAEL HEIZMANN[2]

[1]Intelligent Systems Research Group (ISRG), Karlsruhe University of Applied Sciences, 76133 Karlsruhe, Germany
[2]Institute of Industrial Information Technology (IIIT), Karlsruhe Institute of Technology (KIT), 76187 Karlsruhe, Germany

Corresponding author: Johannes Wetzel (johannes.wetzel@hs-karlsruhe.de)

This work was supported by the German Federal Ministry of Education and Research (BMBF) under Grant 13FH025IX6.

ABSTRACT Privacy-preserving high-quality people detection is a vital computer vision task for various indoor scenarios, e.g. people counting, customer behavior analysis, ambient assisted living or smart homes. In this work a novel approach for people detection in multiple overlapping depth images is proposed. We present a probabilistic framework utilizing a generative scene model to jointly exploit the multi-view image evidence, allowing us to detect people from arbitrary viewpoints. Our approach makes use of mean-field variational inference to not only estimate the maximum a posteriori (MAP) state but to also approximate the posterior probability distribution of people present in the scene. Evaluation shows state-of-the-art results on a novel data set for indoor people detection and tracking in depth images from the top-view with high perspective distortions. Furthermore it can be demonstrated that our approach (compared to the the mono-view setup) successfully exploits the multi-view image evidence and robustly converges in only a few iterations.

INDEX TERMS Depth sensor indoor surveillance, depth sensor networks, generative scene model, joint multi-view person detection, mean-field variational inference, multi-camera person detection, people detection in top-view, vertical top-view pedestrian detection.

I. INTRODUCTION

By virtue of the emergence of low-cost commodity depth sensors, there is an increasing demand for privacy-preserving high-quality people detection in various indoor scenarios, e.g. people counting, customer behavior analysis, public security, ambient assisted living or smart homes. In contrast to classical pedestrian detection approaches, the depth sensors capture the scene from the top-view to minimize occlusions in crowded scenes. However, due to the top-view and the limited mounting height in many indoor scenarios, the resulting field of view of a single depth sensor is quite limited, thus the observable area is rather small. This is an issue in many real-world applications such as customer behavior analysis in a shopping mall or airport. To provide complete detections in a wide-area scenario we therefore employ a multi-view approach. Apart from the increased observable area, there are additional advantages compared to the classical single-view approach. Since a single image does not capture all the details in a 3D scene, considering additional partially overlapping views provides more information about the true scene state. This is especially relevant in situations where people are only partially visible in one camera view due to occlusion or the limited field of view (see Fig. 2). Hence the detection performance (including the reliability of the detection confidence) in the overlapping regions can be improved by the complementary image evidence from multiple views. In particular this is relevant for demanding applications such as emergency detection in an ambient assisted living context.

The general problem of people detection in a multi-camera setup has been widely studied in computer vision literature. However, existing multi-camera people detection approaches mostly focus on outdoor pedestrian detection, capturing pedestrians from profile or frontal view and using monocular video cameras. In contrast, we focus on the task of people detection in multiple overlapping depth images. Due to the vertical top-view, position changes of pedestrians lead to drastically varying appearances, making it very challenging for off-the-shelf data-driven pedestrian detectors without a domain-specific large scale data set. Besides only few methods in the literature take advantage of the full

The associate editor coordinating the review of this manuscript and approving it for publication was Claudio Cusano.

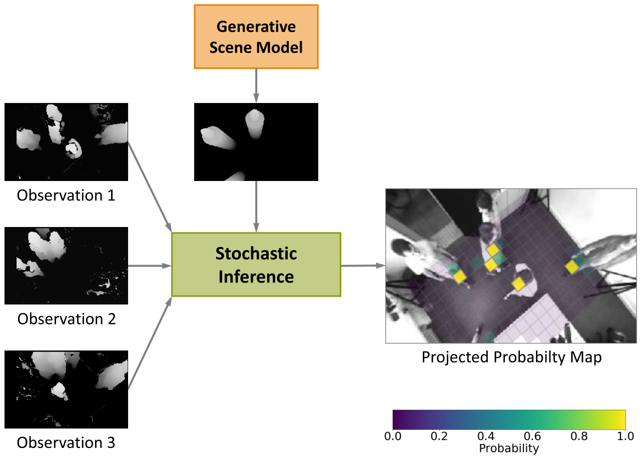


FIGURE 1. Overview of our approach. We use foreground-segmented depth observations from three sensors as input (left) to approximate the marginal probability distribution of people present in the scene (right).

multi-view image evidence from overlapping fields of view, in order to increase the detection performance. To overcome those shortcomings, we propose a novel approach which exploits the multi-view image evidence by (i) employing a generative scene model leading to a viewpoint independent detector without the need of a training data set; (ii) using a probabilistic framework which includes the full multi-view image evidence from all sensors to resolve occlusion as well as measurement noise; and (iii) instead of just estimating the maximum a posteriori (MAP) state, utilizing mean-field variational inference to approximate the posterior distribution of people present in the scene (see Fig. 1). In the evaluation we report state-of-the art results on a novel data set for indoor people detection in multiple top-view depth images.

II. RELATED WORK

Multi-camera people detection has been extensivley studied in the context of video surveillance. The vast majority of the existing approaches is based on multiple monocular video cameras observing an outdoor scene. However, the topic of indoor people detection in multiple depth images, especially in top-view, has not yet been explored in detail. Hence, we will first discuss approaches focusing on pedestrian detection with multiple monocular cameras, and their relation to our approach. In order to restrict our scope, we do not consider methods working across non-overlapping views [1]–[3] but rather focus on methods utilizing overlapping views. For an exhaustive survey of multi-camera people detection and tracking, we refer to [4]–[6]. For the rest of this section, we categorize the relevant literature into approaches utilizing multiple monocular video camera views (RGB-based approaches) and depth images (depth-based approaches).

A. RGB-BASED APPROACHES

Since people detection and tracking in single-camera views have been intensivly studied [7]–[9], many methods accomplish multi-view detection by fusing local detections or local tracklets into a common world coordinate system [10]–[12]. However, since the detection is performed independently for each view, those methods do not take full advantage of the multi-view information, thus making it harder to resolve occlusion and measurement noise. Besides, the vast majority of employed pedestrian detectors is optimized to detect people in frontal or profile view but not in the top-view [13], [14].

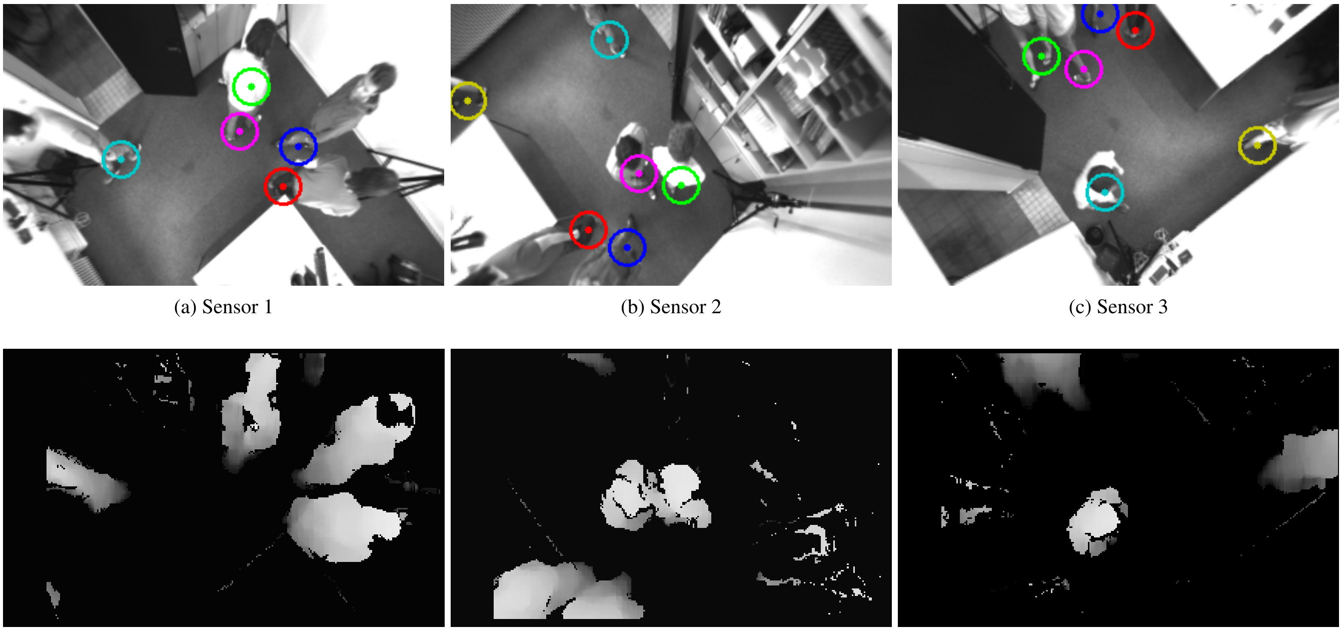
Homography based approaches project local image features from each sensor into a common plane to perform global detection [15]. In [16] a *homographic occupancy constraint* is proposed to handle occlusion and detect people on a common scene plane. Eshel and Moses [17] propose a similar approach, projecting the foreground pixels of all views into a common height plane for head detection. In [18] those approaches are extended by a multi-view Bayesian network in order to avoid false positive detections arising from occlusion artefacts.

Another class of related approaches addresses the problem of multi-camera detection by employing a generative model to jointly take advantage of the image evidence of all available views. Fleuret *et al.* [19] introduce the probabilistic occupancy map (POM). They use foreground-segmented binary images as input and employ a simple person model expressed as a rectangular bounding box to estimate probabilities of occupancy by mean-field variational inference. The method used in our approach is heavily inspired by [19]. Alahi *et al.* [20] re-cast the problem as a linear inverse problem. Other than in [19], a silhouette is proposed as person model. Unlike our approach, both methods utilize only 2D models and fit them to a binary foreground mask.

Baque *et al.* [21] introduce a state-of-the-art end-to-end multi-view people detection architecture. They combine a classical Convolutional Neural Network (CNN) with Conditional Random Fields (CRFs) to resolve ambiguities arising from occlusion. Chavdarova and Fleuret [22] present a CNN architecture to allow for end-to-end multi-view people detection. To overcome the lack of an appropriate multi-view data set, a larger existing monocular pedestrian data set [23] is used. However, due to the lack of extensive labeled data for top-view people detection in depth images, both approaches are insufficient for our use case.

B. DEPTH-BASED APPROACHES

Since people detection in multiple depth images has rarely been studied, we first discuss relevant single-view approaches. The related problem of people counting with a single depth camera from the top-view has been studied in great detail [24]–[27]. In contrast to our proposed method, those approaches focus on integrated systems counting the number of persons crossing a certain virtual line, providing people detection only implicitly and in a rather small area. Recent CNN architectures [28]–[30] are successfully applied to single view depth image people detection leveraging many



(a) Sensor 1
(b) Sensor 2
(c) Sensor 3
(d) Observation 1 (o_1)
(e) Observation 2 (o_2)
(f) Observation 3 (o_3)

FIGURE 2. Example observations from our multi-view setup with six people present in the covered area, marked with unique colors. In view (a) the magenta marked person occludes the green marked person, while in (b) the pair can be clearly separated. In the views (b,c) several people are only partially visible. Notice that for inference only the depth images are used.

labeled images for training. Since in our top-view setup position changes of people lead to drastically varying appearances (compared to the classical frontal or profile view), those approaches need to be re-trained with a domain-specific large-scale data set. Both mentioned classes of methods only provide single-view detection.

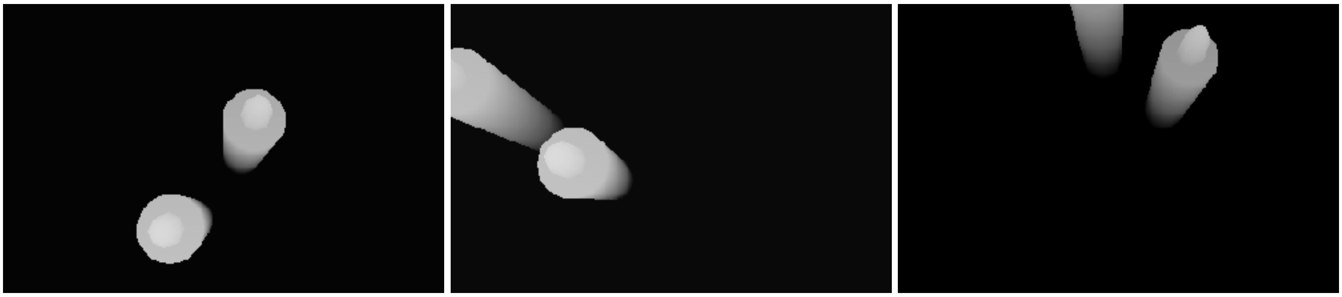
In contrast to the methods mentioned above, only few existing approaches rely on multiple depth images for people detection. Tseng *et al.* [31] present an indoor people detection system based on multiple active sensors in top-view. Their approach is based on a fused virtual top-view depth image, obtained by the point cloud of each sensor. For the detection they employ a hemiellipsoidal head model to take advantage of the discriminative height difference around the head contour of a human. In contrast to our approach, the presented method relies on high-quality depth data. In previous work [32] 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. However, as a consequence of the used optimization method, the number of people in the scene is required a priori, and a sufficiently good initialization is essential. Carraror *et al.* [33] propose an approach for human body pose estimation and tracking in a network of RGB-D sensors. To obtain a global 3D skeleton, CNN-based pose estimation [34] is applied to the RGB images of each single-view. However, due to the single-view detection approach they do not take advantage of the full multi-view information. In contrast to our work, the former approaches [31]–[33] estimate an MAP point estimate but do not provide a probability distribution over people present in the scene.

C. SUMMARY

To summarize, our work is highly inspired by [19]. In contrast, we use depth images as evidence and therefore are able to make use of a more specific generative scene model. We also propose a different strategy to approximate the final mean-field update expectation by making use of geometric scene knowledge and a pre-trained vocabulary. Our generative scene model is similar to our previous work [32]. However, the approach introduced in this work does not hinge on scene-specific a priori knowledge and provides an approximation to the full posterior distribution. In contrast to recent data-driven CNN architectures [22], [28]–[30], [33] our method requires no training data and the detection confidence can be quantified more precisely by approximating the posterior distribution. To the best of our knowledge, variational mean-field inference in combination with a generative scene model has not yet been applied to the problem of people detection in overlapping depth images.

III. APPROACH

The problem we address in this work is the detection of people given multiple overlapping depth images from, but not limited to, the top-view on the scene. The major challenges are (i) the different appearances of people due to the change of viewpoint (see Fig. 2); (ii) occlusions in more crowded scenes and (iii) the measurement noise due to commodity low-resolution depth sensors. To overcome challenge (i), we make



(a) $G_1(\vec{x})$

(b) $G_2(\vec{x})$

(c) $G_3(\vec{x})$

FIGURE 3. Forward model: Synthetic depth images for a given scene configuration $\vec{x}$.

use of a generative scene model (see Fig. 3), formulating the people detection problem as an *analysis-by-synthesis* problem. Challenge (ii) and (iii) are addressed by the proposed probabilistic model (see Sect. III-A), which jointly handles the multi-view information. Furthermore, the mean-field variational inference approach deals with occlusion implicitly and turns our statistical inference problem into a tractable optimization problem, in order to get an approximation of the proposed posterior distribution (see Sect. III-B).

Due to the available depth data, marker-free extrinsic calibration can be achieved in three simple steps: (1) for each sensor $S_1, \ldots, S_C$ the ground floor plane is estimated by a simple plane fit; (2) one arbitrary sensor coordinate system is defined as the common world coordinate system; (3) for each sensor S_c the rigid body transformation to the common world coordinate system is obtained by corresponding natural image features in the overlapping fields of view. For the rest of this paper we define $\mathbf{P}_c$ as the projection matrix for each sensor S_c, which maps a point from the common world coordinate system to the corresponding image coordinates of each sensor.

A. PROBABILISTIC MODEL

Since we assume that the common ground floor plane is known from the initial calibration, we describe the presence of people in the scene in ground floor world coordinates. We discretize the ground floor area into a 2D-grid of n locations. Each location u_i will be assigned a realization x_i of a Bernoulli random variable $X_i \sim \mathcal{B}(p)$, where p denotes the probability of a person present at location u_i. The latent variables are given as the vector $\vec{x} = (x_1, \ldots, x_n)^T \in \{0, 1\}^n$, also referred to as scene configuration. Let $\vec{o} = (o_1, \ldots, o_c)^T$ be the vector of foreground-segmented depth observations at one time step, acquired from depth sensors $S_1 \ldots S_C$. Our joint probability model can be written as

$$p(\vec{o}, \vec{x}) = p(\vec{o}|\vec{x})p(\vec{x}). \tag{1}$$

The likelihood construction is similar to our previous work [32], although we use a discrete grid instead of continuous person locations. To make the likelihood tractable, we assume that the views are conditionally independent for a fixed scene configuration $\vec{x}$. Since we assume that only people are part of the foreground, and that the depth images are robust against illumination changes, this assumption can be justified. Thus, the likelihood factorizes as:

$$p(\vec{o}|\vec{x}) = \prod_{c=1}^{C} p(o_c|\vec{x}). \tag{2}$$

We define the likelihood for one observation by employing a generative forward model $G_c(\vec{x}, \mathbf{P}_c)$, which maps a scene configuration $\vec{x}$ and a given projection matrix $\mathbf{P}_c$ to a synthetic observation (i.e. synthetic depth image) from the perspective of sensor S_c. Therefore, we use a simple, rotationally symmetric 3D person model, consisting of a cylinder for the body and a sphere for the head, see Fig.3. For the sake of simplicity we assume that our given observations suffer from Gaussian noise, yielding an observation likelihood

$$p(o_c|\vec{x}, \sigma) \propto \exp\left(-\frac{1}{2\sigma^2} \|o_c - G_c(\vec{x}, \mathbf{P}_c)\|_2^2\right). \tag{3}$$

Since our generative forward model is not only a function of $\vec{x}$ but also of the projection matrix $\mathbf{P}_c$ we incorporate the physical sensor model in a natural way into our framework, allowing us to detect people from arbitrary viewpoints and to easily integrate a new sensor modality into the network. Applying Bayes' theorem and assuming that the prior factorizes as $p(\vec{x}) = \prod_{i=1}^{n} p(x_i)$, we get the posterior distribution

$$p(\vec{x}|\vec{o}) = \frac{\prod_c p(o_c|\vec{x}) \prod_i p(x_i)}{\sum_{\vec{x}' \in \{0,1\}^n} \prod_c p(o_c|\vec{x}') \prod_i p(x_i')}. \tag{4}$$

B. MEAN-FIELD VARIATIONAL INFERENCE

Because of the dimensionality of the latent scene configuration space $\{0, 1\}^n$, the partition function in (4) is intractable, and we cannot directly compute the posterior distribution. Instead we propose to apply Kullback-Leibler variational inference [35], [36] to approximate the complex distribution $p(\vec{x}|\vec{o})$ by a simpler proxy distribution $q(\vec{x})$. Let $\langle\cdot\rangle_{p(x)}$ be the expectation with respect to a distribution $p(x)$; then the optimization objective can be expressed as

$$\begin{aligned} \hat{q}(\vec{x}) &= \arg\min_q \mathrm{KL}(q(\vec{x}) \,||\, p(\vec{x}|\vec{o})) \\ &= \arg\min_q \langle \log q(\vec{x}) - \log p(\vec{x}|\vec{o}) \rangle_{q(\vec{x})} . \end{aligned} \tag{5}$$

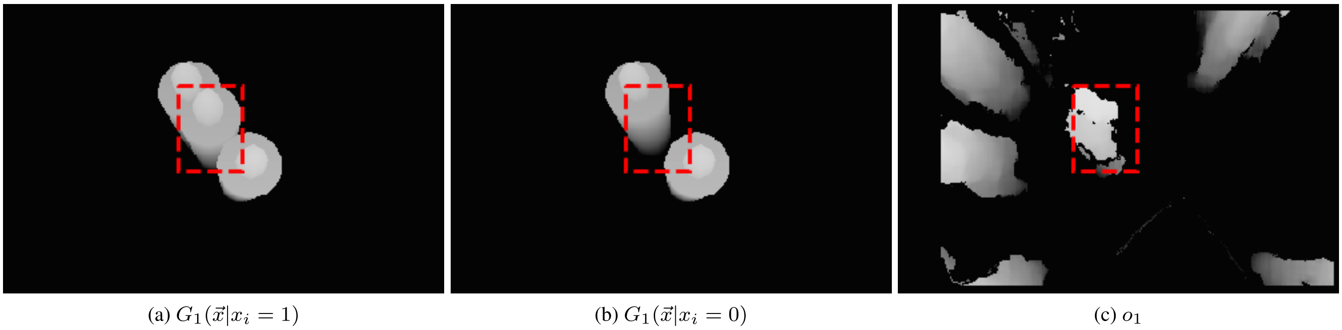


FIGURE 4. The red dashed rectangle illustrates the bounding box $I_1[u_i]$ corresponding to the rendering of a person present at location u_i in sensor 1. The scene configuration $\vec{x}$ shown here is zero for every grid location except for the two neighbors of u_i.

To make the problem computationally tractable, we assume a fully-factorized distribution $q(\vec{x}) = \prod_{i=1}^{n} q_i(x_i)$, known as the *naive mean-field assumption*. Let $q(\vec{x} \setminus x_i)$ denote the mean-field distribution excluding the element x_i, namely $q(\vec{x} \setminus x_i) = \prod_{j=1:j\neq i}^{n} q_j(x_j)$. The general mean field equation, given by

$$q_i(x_i) \propto \exp\left(\langle \log p(\vec{x}|\vec{o})\rangle_{q(\vec{x}\setminus x_i)}\right), \tag{6}$$

updates $q_i(x_i)$ depending on the previous mean-field state $q(\vec{x} \setminus x_i)$. It can be proven that updating $q_i(x_i)$ asynchronously according to (6) will decrease the KL divergence in (5) (see [37, 625 ff.]). Since each x_i is Bernoulli distributed, (6) (for x_i being in state 1) can be written as

$$q_i(x_i = 1) = \frac{1}{Z_i} \exp\left(\langle \log p(\vec{o}, \vec{x}|x_i = 1)\rangle_{q(\vec{x}\setminus x_i)}\right), \tag{7}$$

with the partition function

$$Z_i = \sum_{s\in\{0,1\}} \exp\left(\langle \log p(\vec{o}, \vec{x}|x_i = s)\rangle_{q(\vec{x}\setminus x_i)}\right). \tag{8}$$

Additionally, let $\delta(I_1, I_2) = \frac{1}{2\sigma^2}||I_1 - I_2||_2^2$ be an image distance function, and $\tau_i = \log \frac{1-p(x_i=1)}{p(x_i=1)}$ a function of the prior. Inserting the joint probability distribution defined in (1-4) into (7), and using the relation $\frac{e^x}{e^x+e^y} = \frac{1}{1+e^{y-x}}$, the final asynchronous update of the probability $q_i(x_i = 1)$ is a sigmoid function given as

$$q_i(x_i = 1) = \left[1 + \exp\left(\tau_i + \sum_{c=1}^{C} E_{c,i}\right)\right]^{-1}, \tag{9}$$

with the expectation

$$E_{c,i} = \langle \delta(o_c, G_c(\vec{x}|x_i{=}1)) - \delta(o_c, G_c(\vec{x}|x_i{=}0))\rangle_{q(\vec{x}\setminus x_i)} \,. \tag{10}$$

Notice that $G_c(\vec{x}|x_i = 1)$ maps a scene configuration $\vec{x}$ to a synthetic depth image in the perspective of sensor S_c with x_i forced to 1 (see Fig. 4). Following the argument given in [19], one can see how occlusion is handled in an implicit way: If the forward-model projection of a person located at u_i is occluded by a projection of a person with a high probability of occupancy, the value of x_i does not affect the image distance $\delta(o_c, G_c(\vec{x}|x_i = s))$. Thus, the expectation $E_{c,i}$ in (10) converges to zero.

C. APPROXIMATE MEAN-FIELD UPDATE

Still (10) is intractable due to the expectation $\langle\cdot\rangle_{q(\vec{x}\setminus x_i)}$, which implies an iteration over all scene configurations. We approximate the expected value by considering only a relevant subset of scene configurations. Therefore, we exploit the fact that the difference

$$\delta(o_c, G_c(\vec{x}|x_i = 1)) - \delta(o_c, G_c(\vec{x}|x_i = 0)) \tag{11}$$

only depends on the pixels belonging to the silhouette of the projection of the 3D model at location u_i (see Fig. 4).

For a simpler and faster implementation, we do not work on the exact silhouettes but on the corresponding rectangular bounding boxes, given as $I_c[u_i]$. Thus only those scene configurations, for which the pixel values inside the bounding box $I_c[u_i]$ of the generated image $G_c(\vec{x})$ are effected, need to be evaluated for the expectation $E_{c,i}$ in (10). We assume that only the projections of the direct eight neighbors of a grid location u_i intersect with the bounding box $I_c[u_i]$. For our top-view setup this is a valid assumption; however, for a frontal view setup, a more sophisticated approximation would be preferable. Consequently, we can approximate the expectation $E_{c,i}$ in (10) by the reduced neighborhood scene configuration $\tilde{\vec{x}}_i \in \{0, 1\}^8$. Since the local neighborhood (including x_i) allows only $2^9 = 512$ possible scene configurations, we can effectively approximate the expectation.

Instead of the image distance $\delta(\cdot, \cdot)$, derived from our probabilistic model, we introduce a weighted asymmetric image similarity $\delta_{\text{asym}}(o, g)$ between a foreground segmented observation o and a generated image g. Since there is no need to compute the derivative of the distance function we replace the squared L2-norm by the more robust L1-norm. Let $M : \mathbb{R}^{W\times H} \mapsto \{0, 1\}^{W\times H}$ be a threshold function which maps an image to its binary foreground mask, $\overline{M}(i) = 1 - M(i)$ its inverse and $\odot$ the hadarmard product between two images. The asymmetric image similarity is given as

$$\begin{aligned}\delta_{\text{asym}}(o, g) = \Big(\alpha \left\|o \odot \overline{M}(g)\right\|_1 + (2-\alpha)\left\|g \odot \overline{M}(o)\right\|_1 \\ + \|(o-g) \odot M(o) \odot M(g)\|_1\Big)\,,\end{aligned} \tag{12}$$

with the design parameter $\alpha \in [0, 2]$. For $\alpha = 1$ the image similarity $\delta_{\text{asym}}(o, g)$ is identical to the L1-norm $\|o - g\|_1$. For $\alpha > 1$ observed depth pixels which are not explained by the generative scene model will be penalized stronger. Let further

$$\delta_{x_i=s} = \frac{1}{2\sigma^2}\delta_{\text{asym}}(o_c[u_i], G_c(\tilde{\tilde{x}}_i|x_i = s)[u_i]) \tag{13}$$

be the image similarity restricted to the cropped image region $I_c[u_i]$. Then the approximated expectation can be written as

$$\tilde{E}_{c,i} = \frac{1}{|I_c[u_i]|}\left\langle \delta_{x_i=1} - \delta_{x_i=0}\right\rangle_{q(\tilde{\tilde{x}}_i)} . \tag{14}$$

Additionally, we normalize the expectation with respect to the size of the image slice $|I_c[u_i]|$, to account for the viewpoint dependent size of a bounding box. In order to efficiently compute (14), we propose to pre-build for each u_i a vocabulary of image sections $I_c[u_i]$ for all 512 possible scene configurations $\tilde{x}_i$.

The final mean-field updates can be executed asynchronously or synchronously. In an asynchronous mean-field update iteration, the individual $q_i(x_i)$'s are updated sequentially, whereas, in a synchronous update iteration, all the $q_i(x_i)$ are updated simultaneously, using the same previous mean-field state $q(\vec{x})$. While asynchronous update provides theoretical convergence (see III-B), synchronous mean-field updates can be easily parallelized. For the optimization we use coordinate-ascent variational inference (CAVI) [35]. Hence, the probability for each $\hat{q}_i(x_i)$ will asynchronously be updated with respect to the previous mean-field state $q(\vec{x})$ according to the final update equation

$$\hat{q}_i(x_i = 1) = \left[1 + \exp\left(\tau_i + \sum_{c=1}^{C}\tilde{E}_{c,i}\right)\right]^{-1} . \tag{15}$$

IV. EVALUATION

A. DATA SET

To the best of our knowledge, currently no publicly available data set that covers the scenario of top-view people detection using multiple depth sensors with overlapping fields of view exists. Therefore, we introduce a novel data set to compare our approach with state-of-the art multi camera people detection approaches. The data set contains footage from an indoor office scene and is recorded from three low-resolution commodity stereo-vision-based depth sensors, covering a variety of constellations (see Fig. 7). The sensors have a top-view on the scene, are mounted at a height of three meters, and have fields of view with a significant joint overlap (see Fig. 2). They cover a visible area of approximately $20\,m^2$ with up to six individual people present in the scene, entering and leaving the visible area multiple times. The data set consists of 2200 annotated frames, captured with a resolution of 376 × 240 pixel each, providing raw rectified stereo image pairs as well as disparity maps obtained by block matching. In total we annotated the ground floor locations of 10435 targets. Additionally, we associated each detection with a track to allow for full detection and tracking evaluation. For the reproducibility of our results the data set will be made publicly available.

B. QUANTITATIVE ANALYSIS

For the evaluation of our approach we use a ground floor grid with 15 × 12 grid points, corresponding to a horizontal and vertical distance of 33 *cm* between adjacent grid points. As input observations we use foreground-segmented depth images, obtained by static background subtraction. Notice that we only focus on frame-by-frame detection, however the outcome of our approach could serve as input for tracking-by-detection post-processing.

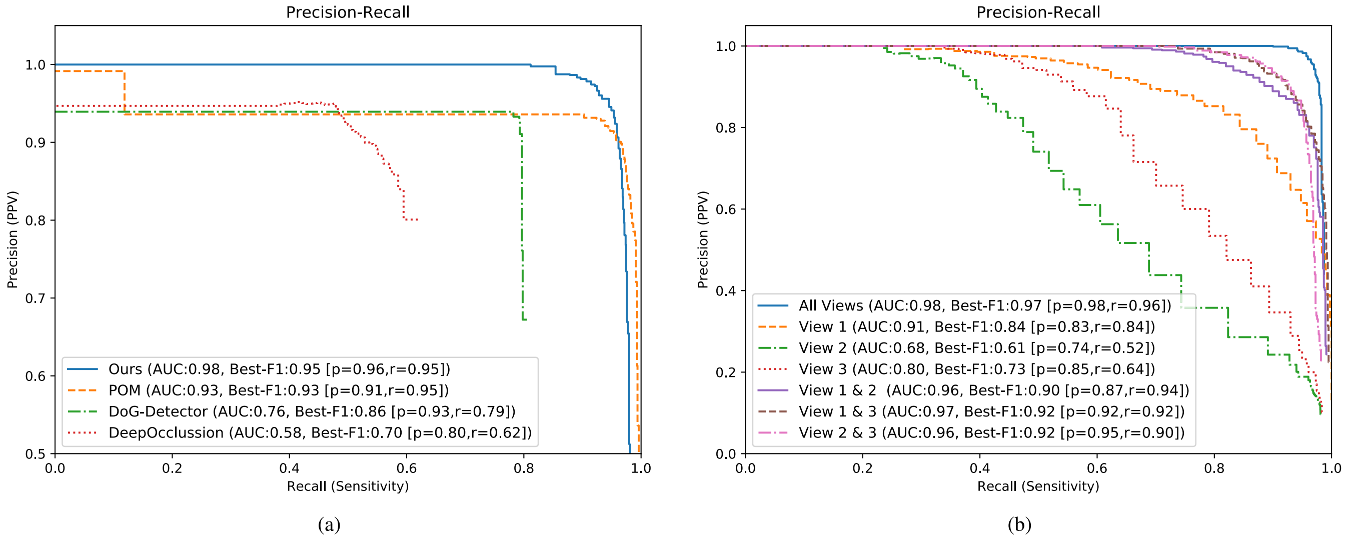
We have noticed that our approach is quite sensitive to the initial marginal probabilities $q_i^{init}(x_i)$. If the initial occupancy probability is too small, the expectation in (10) will inordinately favor scene configurations with only one person present; thus, occlusion is not taken into account in the first iteration. We therefore initialize each mean-field node with a prior of $q_i^{init}(x_i) = p(x_i) = 0.5$ by default. The design parameter of the asymmetric image similarity $\delta_{asym}(\cdot, \cdot)$ (see (12)) is set to $\alpha = 1.25$, to penalize unexplained observations stronger. Fig. 6 depicts the impact of α on the precision-recall performance. The standard deviation of the measurement noise σ is set to a default value of 2 *cm*.

For the quantitative evaluation, a detection is assumed to be a true positive if it is in a radius of 30 *cm* of the ground truth. We show the performance of our approach based on the precision-recall curves in Fig. 5, 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 is given as $F_1 = (2 \times \text{precision} \times \text{recall})/(\text{precision} + \text{recall})$.

We compare our approach with state-of-the-art monocular multi-view approaches. As a baseline on the given depth observations we introduce a difference of Gaussian (DoG) based blob detector. The methods to be compared are:

- **POM** [19] works on binary input observations. For a fair comparison we use the same depth based foreground segmentation as in our approach. Also the grid layout and the camera calibrations are identical to our setup.
- **Deep Occlussion** [21] is the current state-of-the art end-to-end architecture for multi-view person detection. Due to the lack of a large data set we use the available pre-trained model without any further supervision. As input we stack the given gray scale observations to a three channel image to be compatible with the RGB architecture.
- **DoG-Detector** As a baseline on the given depth data we apply difference of Gaussian blob detection on the foreground segmented depth images of each sensor independently and project the resulting detections onto the common world ground plane. The final detections on the ground plane are obtained by proximity clustering.

Fig. 5a depicts the performance of the examined approaches over all frames and views. The results show that without any further supervision the given data set is very challenging for deep learning architectures such as Deep Occlusion [21]. Due to the vertical top-view, the appearances



(a)

(b)

FIGURE 5. Precision-recall curves showing the performance of our approach. In (a) the precision-recall performance (precision range $[0.5, 1]$) over all frames and views is plotted. For (b), only the people visible in all three views are taken into account.

TABLE 1. Performance of evaluated approaches.

	AUC	F1-Score	Precision	Recall
Ours	**0.98**	**0.95**	**0.96**	**0.95**
POM [19]	0.93	0.93	0.91	**0.95**
DoG-Detector	0.76	0.86	0.93	0.79
Deep Occlussion [21]	0.58	0.70	0.80	0.62

of people is drastically different compared to the classical profile view. The results of the DoG-Detector indicate that, even when considering proximity clustered results from all three views, naive blob-based single-view detectors are not competitive compared to the more sophisticated multi-view approaches in our scenario. Although POM [19] achieves remarkable performance in our setting, our approach outperforms POM in terms of precision, resulting in a better area under the curve value (AUC) as well as better F1-Score (see Table. 1).

In order to show how our probabilistic model exploits the multi-view evidence given by all three sensors, we evaluated the performance of our approach for all different combinations of sensor views contributing to the solution. For a fair comparison, we take only those people into account that are visible from all three sensors (see Fig. 2 for the fields of view of the sensors). Fig. 5b depicts how using the multi-view information increases the detection performance. In the mono-view case, View 2 and View 3 by themselves do not perform well with F1-Scores of 0.61 and 0.73, respectively. However, combining the image evidence of View 2 and View 3 leads to a drastic performance increase, as evidenced by a best F1-Score of 0.92. Even View 1 achieves comparable good performance due to the general viewpoint, using the image evidence from all three sensors clearly outperforms all other view combinations.

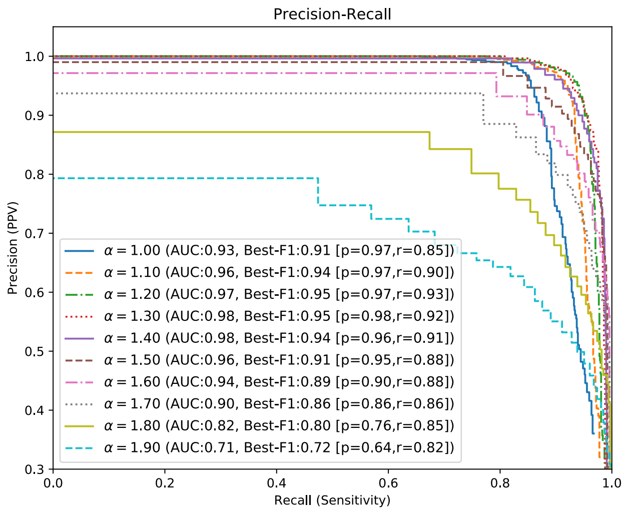


FIGURE 6. Precision-recall curves for different values of the asymmetric image similarity parameter α.

On a single CPU core,[1] our non-optimized Python implementation needs approximately 800 *ms* per frame. Although real-time performance is not reached yet, there are plenty of optimization options, such as parallel mean-field updates, or taking advantage of GPUs.

C. QUALITATIVE ANALYSIS

Fig. 7 shows exemplary mean-field optimization results. The given samples illustrate that our approach is able to resolve challenging scenarios, suffering from occlusion and measurement noise, by making use of the full multi-view image evidence. Fig. 7c shows a typical false negative error on the

[1] Intel Core-i7@2.9Ghz

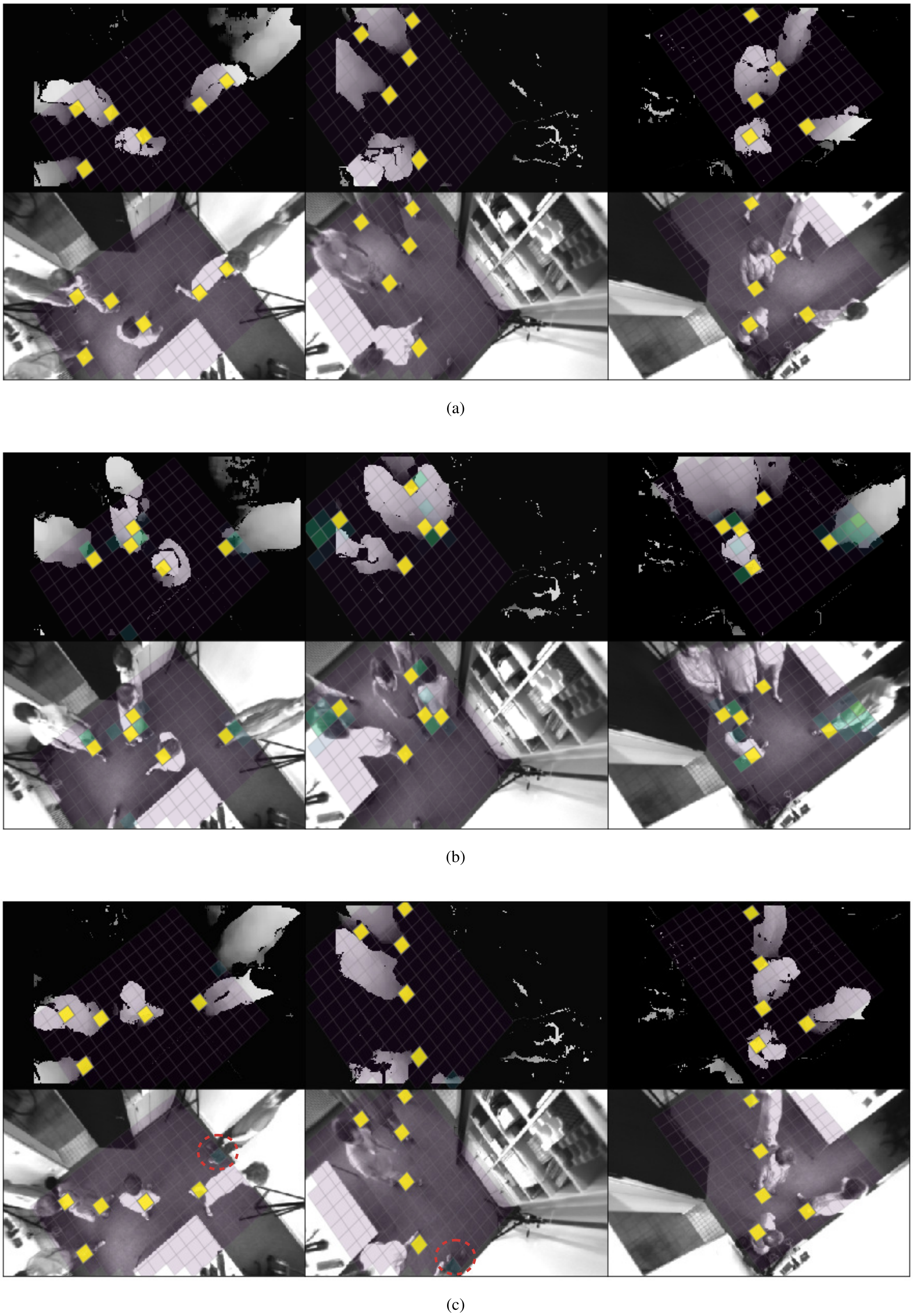
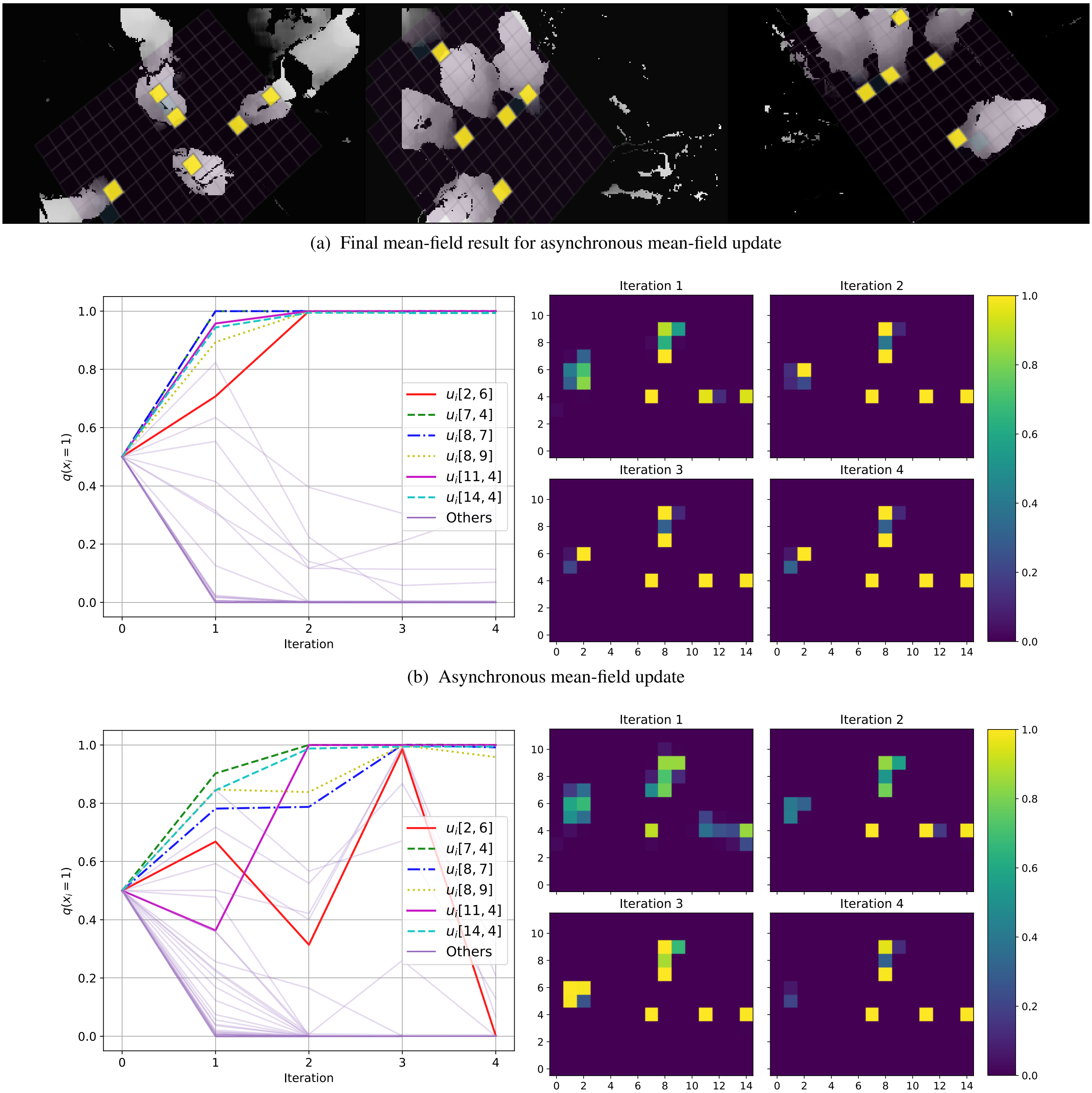


FIGURE 7. Exemplary mean-field optimization result $\hat{q}(\bar{x})$ after five iterations. The final marginal probability map is projected onto the ground floor, where purple correspond to a probability of zero and yellow to one respectively. (a) Shows an estimation of the marginal probability distribution $q(\bar{x})$ with clear peaks at grid locations occupied by a person. (b) Includes some uncertainty around the peaks of the distribution. (c) Includes a typical false negative, marked with a red dashed circle.



(a) Final mean-field result for asynchronous mean-field update

(b) Asynchronous mean-field update

(c) Synchronous mean-field update

FIGURE 8. Evolution of asynchronous and synchronous mean-field updates. In the left-hand plots of (b) and (c), every path corresponds to the probability evolution of one $q_i(x_i)$. The probability evolution of six grid locations of interest are plotted in unique colors, the others are plotted in purple. The right-hand plots show the same process illustrated as probability maps for the first four iterations.

image border, which is the dominant error class occurring in the data set. Due to the stereo vision based sensors, the depth information is more noisy on the image border, eventually leading to an insufficient fit of the 3D model. To overcome this limitation, a richer probabilistic sensor model which takes systematically varying noise into account could be employed.

In Fig. 8, the mean-field optimization is illustrated for one exemplary frame, for both the asynchronous and the synchronous update strategy. Fig. 8c depicts a general disadvantage of synchronous mean-field updates. The simultaneous optimization potentially leads to oscillating marginal probabilities of adjacent grid locations. In Fig. 9 it is shown that asynchronous mean-field optimization converges after

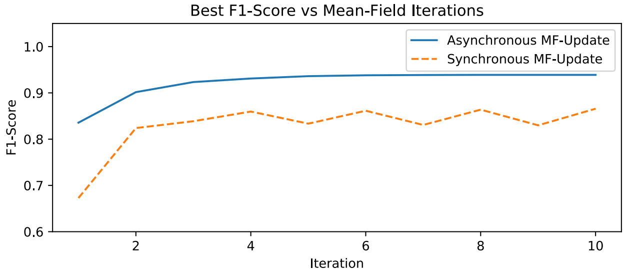


FIGURE 9. Comparison of asynchronous and synchronous mean-field updates.

only few iterations whereas synchronous mean-field update suffers from the oscillating effects mentioned above.

V. CONCLUSION

In this work we have presented a novel approach for probabilistic people detection in multiple overlapping depth images. Our main contribution is the use of mean-field variational inference in combination with a generative scene model to jointly exploit the multi-view information in order to approximate the marginal probability distribution of people present in the scene. Our experiments have shown state-of-the-art results on a novel data set for indoor people detection in overlapping depth images from the top-view. We have demonstrate that our approach achieves strong detection performance, outperforming state-of-the-art monocular multi-view people detection methods. We were also able to show that using multi-view image evidence increases the detection performance significantly compared to a single-view.

Future work will focus on incorporating temporal information into our probabilistic model in order to provide joint probabilistic detection and tracking.

REFERENCES

[1] L. Beyer, S. Breuers, V. Kurin, and B. Leibe, "Towards a principled integration of multi-camera re-identification and tracking through optimal Bayes filters," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1444–1453.

[2] A. Rahimi, B. Dunagan, and T. Darrell, "Simultaneous calibration and tracking with a network of non-overlapping sensors," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Nov. 2004, pp. 187–194.

[3] E. Ristani and C. Tomasi, "Features for multi-target multi-camera tracking and re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6036–6046.

[4] 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, no. 1, p. 43, Dec. 2017.

[5] R. Iguernaissi, D. Merad, K. Aziz, and P. Drap, "People tracking in multi-camera systems: A review," *Multimedia Tools Appl.*, vol. 78, no. 8, pp. 10773–10793, Apr. 2019.

[6] X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern Recognit. Lett.*, vol. 34, no. 1, pp. 3–19, Jan. 2013.

[7] H. Kieritz, S. Becker, W. Hubner, and M. Arens, "Online multi-person tracking using integral channel features," in *Proc. 13th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2016, pp. 122–130.

[8] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast R-CNN for pedestrian detection," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 985–996, Oct. 2018.

[9] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Towards reaching human performance in pedestrian detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 973–986, Apr. 2018.

[10] N. Anjum and A. Cavallaro, "Trajectory association and fusion across partially overlapping cameras," in *Proc. IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Sep. 2009, pp. 201–206.

[11] G. Kayumbi, N. Anjum, and A. Cavallaro, "Global trajectory reconstruction from distributed visual sensors," in *Proc. 2nd ACM/IEEE Int. Conf. Distrib. Smart Cameras (ICDSC)*, Sep. 2008, pp. 1–8.

[12] Y. Xu, X. Liu, Y. Liu, and S.-C. Zhu, "Multi-view people tracking via hierarchical trajectory composition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4256–4265.

[13] I. Ahmed and A. Adnan, "A robust algorithm for detecting people in overhead views," *Cluster Comput.*, vol. 21, no. 1, pp. 633–654, Mar. 2018.

[14] 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 Workshop*, 2017, pp. 1–9.

[15] A. Sankaranarayanan, A. Veeraraghavan, and R. Chellappa, "Object detection, tracking and recognition for multiple smart cameras," *Proc. IEEE*, vol. 96, no. 10, pp. 1606–1624, Oct. 2008.

[16] S. Khan and M. Shah, "Tracking multiple occluding people by localizing on multiple scene planes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 505–519, Mar. 2009.

[17] R. Eshel and Y. Moses, "Homography based multiple camera detection and tracking of people in a dense crowd," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[18] P. Peng, Y. Tian, Y. Wang, J. Li, and T. Huang, "Robust multiple cameras pedestrian detection with multi-view Bayesian network," *Pattern Recognit.*, vol. 48, no. 5, pp. 1760–1772, May 2015.

[19] 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, Feb. 2008.

[20] A. Alahi, L. Jacques, Y. Boursier, and P. Vandergheynst, "Sparsity driven people localization with a heterogeneous network of cameras," *J. Math. Imag. Vis.*, vol. 41, nos. 1–2, pp. 39–58, Sep. 2011.

[21] P. Baque, F. Fleuret, and P. Fua, "Deep occlusion reasoning for multi-camera multi-target detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 271–279.

[22] T. Chavdarova and F. Fleuret, "Deep multi-camera people detection," in *Proc. 16th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2017, pp. 848–853.

[23] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 304–311.

[24] L. D. Pizzo, P. Foggia, A. Greco, G. Percannella, and M. Vento, "Counting people by RGB or depth overhead cameras," *Pattern Recognit. Lett.*, vol. 81, pp. 41–50, Oct. 2016.

[25] S. Sun, N. Akhtar, H. Song, C. Zhang, J. Li, and A. Mian, "Benchmark data and method for real-time people counting in cluttered scenes using depth sensors," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 10, pp. 3599–3612, Oct. 2019.

[26] V. Carletti, L. D. Pizzo, G. Percannella, and M. Vento, "An efficient and effective method for people detection from top-view depth cameras," in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2017, pp. 1–6.

[27] X. Zhang, J. Yan, S. Feng, Z. Lei, D. Yi, and S. Z. Li, "Water filling: Unsupervised people counting via vertical Kinect sensor," in *Proc. IEEE 9th Int. Conf. Adv. Video Signal-Based Surveill.*, Sep. 2012, pp. 215–220.

[28] X. Liu, L. Mei, D. Yang, J. Lai, and X. Xie, "Feature visualization based stacked convolutional neural network for human body detection in a depth image," in *Pattern Recognition and Computer Vision* (Lecture Notes in Computer Science), vol. 11257. Cham, Switzerland: Springer, 2018, pp. 87–98. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-03335-4_8

[29] 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. Multimedia*, vol. 20, no. 9, pp. 2249–2261, Sep. 2018.

[30] H. Li, J. Liu, G. Zhang, Y. Gao, and Y. Wu, "Multi-glimpse LSTM with color-depth feature fusion for human detection," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 905–909.

[31] 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/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2014, pp. 4077–4082.

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

[33] M. Carraro, M. Munaro, J. Burke, and E. Menegatti, "Real-time markerless multi-person 3D pose estimation in RGB-depth camera networks," in *Intelligent Autonomous Systems* (Advances in Intelligent Systems and Computing), vol. 867. Cham, Switzerland: Springer, Jun. 2019, pp. 534–545.

[34] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4724–4732.

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

[36] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *FNT Mach. Learn.*, vol. 1, nos. 1–2, pp. 1–305, 2007.

[37] D. Barber, *Bayesian Reasoning and Machine Learning*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

JOHANNES WETZEL received the B.S. and M.S. degrees in computer science from the Karlsruhe University of Applied Sciences, Germany, in 2012 and 2014, respectively. He is currently pursuing the Ph.D. degree in computer vision with the Karlsruhe Institute of Technology (KIT), Germany.

From 2014 to 2017, he was a Computer Vision Research and Development Engineer with Vitracom AG, Karlsruhe. Since 2017, he has been with the Intelligent Systems Research Group (ISRG), Karlsruhe University of Applied Sciences, Germany. His research interests include probabilistic modeling and inference, computer vision, machine learning, and its applications in industry.

ASTRID LAUBENHEIMER received the Diploma degree in mathematics and the Ph.D. degree in 3D model-based computer vision from the University of Karlsruhe, Germany, in 1998 and 2004, respectively.

From 2004 to 2006, she was a Postdoctoral Assistant at the Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (IOSB), Karlsruhe, Germany. From 2006 to 2009, she was the Head of the Research Group Image-Based Real-Time Systems, Fraunhofer IOSB. Since 2009, she has been a Full Professor in applied computer science at the University of Applied Sciences in Karlsruhe (HSKA), where she is a member and a spokeswoman of the Intelligent Systems Research Group (ISRG). Her research interests include probabilistic modeling and inference, machine learning, and mainly with the applications in computer vision and 3D modeling.

MICHAEL HEIZMANN received the M.S. degree in mechanical engineering and the Ph.D. degree in automated visual inspection from the University of Karlsruhe, Germany, in 1998 and 2004, respectively.

From 2004 to 2009, he was a Postdoctoral Research Assistant with the Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (IOSB), Karlsruhe, Germany. From 2009 to 2016, he was the Head of the department Systems for Measurement, Control and Diagnosis (MRD), Fraunhofer IOSB. From 2014 to 2016, he was an Additional Professor of mechatronic systems at the University of Applied Sciences in Karlsruhe. Since 2016, he has been a Full Professor of mechatronic measurement systems and the Director of the Institute of Industrial Information Technology (IIIT), Karlsruhe Institute of Technology (KIT). His research interests include measurement and automation technology, machine vision and image processing, and image and information fusion.