People Detection in a Depth Sensor Network via Multi-View CNNs trained on Synthetic Data

Johannes Wetzel*, Samuel Zeitvogel*, Astrid Laubenheimer* and Michael Heizmann†
* *Intelligent Systems Research Group (ISRG), Karlsruhe University of Applied Sciences, Karlsruhe, Germany*
{johannes.wetzel,samuel.zeitvogel,astrid.laubenheimer}@hs-karlsruhe.de
† *Institute of Industrial Information Technology (IIIT), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany*
michael.heizmann@kit.edu

***Abstract*—In this work an approach for wide-area indoor people detection with a network of depth sensors is presented. We propose an end-to-end multi-view deep learning architecture which takes three foreground segmented overlapping depth images as input and predicts the marginal probability distribution of people present in the scene. In contrast to classical data-driven approaches our method does not make use of any real image data for training but uses a randomized generative scene model to generate synthetic depth images which are used to train our proposed deep learning architecture. The evaluation shows promising results on a publicly available data set.**

***Index Terms*—multi-view person detection; network of depth cameras; top-view people detection; synthetic depth images; multi-view CNN architecture**

I. Introduction

Wide-area people detection is an important prerequisite for various indoor applications, e.g. people counting, customer behavior analysis, public security or smart homes. In this work we address the problem of indoor people detection with a network of depth sensors. Typically, the depth sensors capture the scene from the vertical top-view to reduce occlusions in crowded scenes. As a consequence of the top-view, the appearances of pedestrians varies drastically, making it very challenging for off-the-shelf data-driven pedestrian detectors without a domain-specific large scale data set. To overcome the lack of a large-scale data-set for depth image people detection we present a randomized generative scene model to generate a synthetic training data set of arbitrary size. Moreover, we propose a multi-view CNN architecture which is only trained by the randomized synthetic depth images to predict the marginal probability distribution of people present in the scene (see Fig. 1). In the evaluation we compare our method with state-of-the-art multi-view people detection methods.

While the related task of multi-view people detection and tracking with monocular video cameras has been studied in great detail [1]–[4] only a few approaches in the literature rely on a network of depth sensors for people detection. In the following we discuss existing approaches based on multiple depth images and refer to [5] for a broader discussion on related single-view or monocular multi-camera methods.

Tseng *et al.* [6] 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, back projected from the 3D points obtained by 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. Carraror *et al.* [7] 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 is applied to the RGB images of each single-view. In previous work [8] 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 [5] based on a discrete scene configuration space. For stochastic inference a variational mean-field approximation is used to jointly exploit the multi-view information in order to estimate the marginal probability distribution of people present in the scene.

For inference the mentioned probabilistic methods [5], [8] perform iterative optimization, which is computationally intensive and potentially prone to local minima in the optimization objective. In contrast, we propose an end-to-end CNN architecture, which is demanding at training time, but once the network is trained, inference can be obtained by a single deterministic forward pass. Due to the lack of a domain-specific large scale data set we extend the generative scene model proposed in [5] to generate randomized synthetic training data. In contrast to classical data-driven approaches the proposed multi-view CNN architecture is only trained with synthetic depth images and does not rely on any real training data.

II. Approach

We propose an end-to-end CNN architecture which takes three foreground segmented depth images as input and predicts the marginal probabilities of people present in the scene (see Fig. 1). For this work we assume that the common ground floor plane is known from the initial calibration, hence the presence of people in the scene is expressed in ground floor world coordinates. We discretize the ground floor area into a 15×12 grid of $n = 180$ locations. Each location u_i is assigned to a realization x_i of a Bernoulli random

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

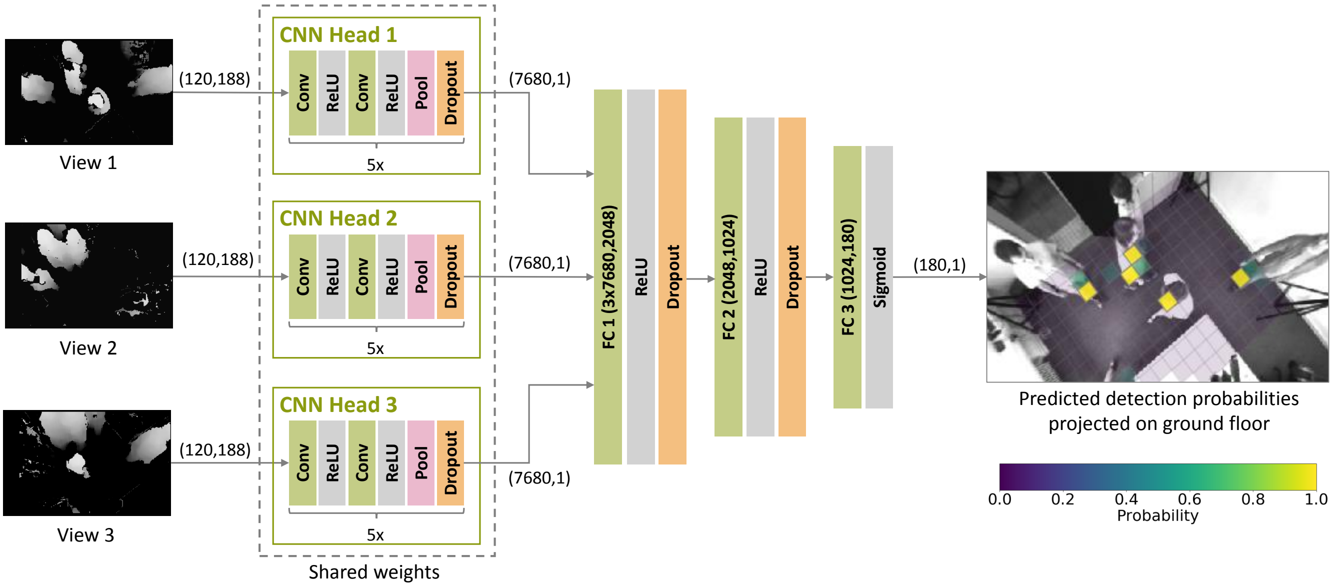

Fig. 1. Overview of our proposed CNN multi-view architecture. Each input depth image serves as input for a CNN module. The output of the last fully connected layer predicts the marginal probability distribution of people present in the scene.

variable $X_i \sim \mathcal{B}(p)$, where p denotes the probability of a person present at location u_i. A scene configuration is given as the vector $\vec{x} = (x_1, \ldots, x_n)^T \in \{0,1\}^n$. Let further $\vec{o} = (o_1, \ldots, o_c)^T$ be the vector of foreground-segmented depth observations at one time step. The objective of our end-to-end approach is to approximate the distribution

$$p(\vec{x}|\vec{o}) = \prod_{i=1}^{n} p(x_i|\vec{o}) \qquad (1)$$

with $p(x_i|\vec{o})$ being the marginal probability of a person present at ground floor location u_i given the observations (see Fig. 1). To approximate (1) we propose a multi-view CNN architecture which jointly exploits the depth observations from three sensors (see Sect. II-A) by leveraging a generative scene model (see Sect. II-B) for randomized synthetic training data generation.

A. End-to-End CNN Architecture

The proposed architecture is depicted in Fig. 1. We observed only a slight drop in performance when the input depth images are down scaled by a factor of 0.5, thus we use input depth images of size 188 × 120 for each individual CNN-head. To generalize over the visual features, weights are shared across the input CNN-heads. The resulting feature maps of each CNN-head are concatenated and fed into a fully connected layer in order to learn correlations between the individual views. Each CNN-head is built of five blocks sharing the same structure. After each block a dropout layer with $p_{CNN} = 0.25$ is applied. The parameters of the CNN layers are given in Tab. I. The three resulting feature vectors are concatenated and used as input for the first fully connected layer `FC1`. After the first two fully connected layers, dropout with $p_{FC} = 0.5$ is used to prevent over-fitting. The final fully connected layer `FC3` is followed by a sigmoid function and predicts the desired marginal probabilities of people present in the scene (1). To train the network we formulated the estimation of the desired marginal probabilities as a binary classification problem, thus using the binary cross-entropy loss

CNN block	Layer type	Filters	Kernel Size
1	Conv (1,*)	32	5 × 5
2	Conv (2,*)	64	3 × 3
3	Conv (3,*)	128	3 × 3
4	Conv (4,*)	256	3 × 3
5	Conv (5,*)	512	3 × 3
1-5	Max Pool	*	2 × 2

TABLE I
PARAMETERS OF CNN HEADS.

$$l_{bce} = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(\hat{p}(x_i)) + (1-y_i)\log(1-\hat{p}(x_i)), \quad (2)$$

with $\vec{y} = (y_1, \ldots, y_n)^T \in \{0,1\}^n$ being the ground truth scene configuration and $\hat{p}(x_i)$ being the predicted probability of a person present at cell u_i.

B. Randomized Generative Scene Model

The generative scene model used in this work is based on the model proposed in previous work [5]. The basic model is built on a static rotationally symmetric 3D person model, consisting of a cylinder for the body and a sphere for the head, see Fig. 2(a). To generate synthetic samples the person model is placed accordingly to the scene configuration $\vec{x}$ on the world ground floor and synthetic depth images are rendered in the perspective of each sensor. We extend the static person model by introducing a parameterized person model to express different shapes of persons in the scene. To achieve randomization we treat the parameters as random

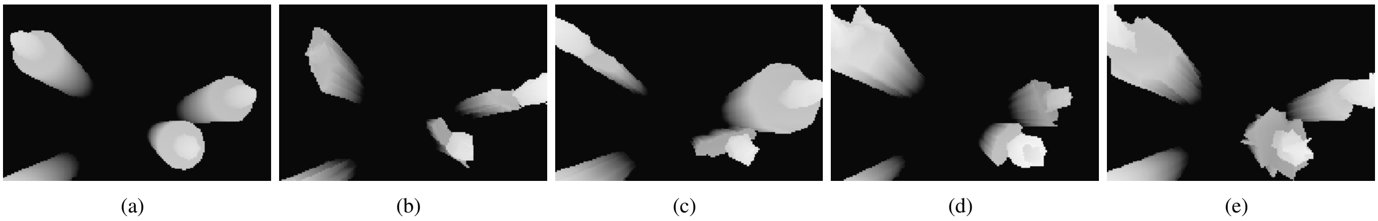

Fig. 2. Synthethic depth images generated from our scene model for one specific scene configuration $\vec{x}$ in sensor view one. Fig (a) shows the synthetic depth image based on our static person model as used in [5]. Fig. (b)-(e) shows four independently drawn samples from our proposed randomized person model for the same scene configuration $\vec{x}$.

variables. Each person model is defined by a set of vertices $\mathcal{V} = \{\vec{v}_1, \dots, \vec{v}_m\}$ with $\vec{v} = (x, y, z)^T$ and a set of faces $\mathcal{F}$ where each face is given by a triple of vertices. We split the set of vertices of one person in two subsets, one for the sphere and one for the cylinder, thus $\mathcal{V} = \mathcal{V}_{cyl} \cup \mathcal{V}_{sph}$, to be able to apply transformations independently on the two geometric primitives. As world coordinate system we define the z-axis perpendicular to the ground plane (xy-plane with $z = 0$), representing the height over ground. It is assumed that a person mesh is initially centered in the xy-plane with the foot point at $z = 0$. To take variants in pose and shape into account, we suggest three principle degrees of freedom: (i) deforming the body of a person (circular cylinder) to an elliptic cylinder to get a variety of rotationally asymmetric shapes; (ii) rotating the person model around the z-axis to model the body orientation; (iii) resize the height of a person. To apply these variants, we define the parameterized transformation

$$f(\vec{v}; s_x, s_y, s_z, \alpha) = \mathbf{R}(\alpha) \cdot \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & s_z \end{bmatrix} \cdot \vec{v}, \qquad (3)$$

which applies non-uniform scaling followed by a rotation around the z-axis with angle α on a single vertex. To generate a single instance we apply the transformation $f(\vec{v}; S_x^c, S_y^c, S_z^c, \alpha)$ to all vertices in set $\mathcal{V}_{cyl}$ and $f(\vec{v}; 1, 1, S_z^s, 0)$ to all vertices in the set $\mathcal{V}_{sph}$ respectively. The parameters $S_x^c, S_y^c, S_z^c, S_z^s, \alpha$ are considered to be uniformly distributed random variables (see. Algorithm 1). To get more variations in shape we add independent Gaussian noise to the x, y, z-components of each vertex $\vec{v} \in \mathcal{V}$. A detailed description of the sampling process and the assumed parameter distributions are given in Algorithm 1. Fig. 2(b)-2(e) show exemplary sampled synthetic depth images for a scene configuration $\vec{x}$.

Algorithm 1. Randomized generation of synthetic depth images.

```
procedure SAMPLEFROMGENERATIVEMODEL
    $\mathcal{V}_{cyl}, \mathcal{V}_{sph}, \mathcal{F} \leftarrow$ init()          ▷ init with default model
    $h \sim \mathcal{U}(2, 6)$          ▷ drawn number of expected persons
    $\vec{x} \sim \mathcal{B}(1/h)$          ▷ draw scene configuration
    for all $x_i = 1$ do          ▷ iterate over cells with a person
        $S_x^c, S_y^c \sim \mathcal{U}(0.5, 1.5)$
        $S_z^c \sim \mathcal{U}(0.85, 1.15)$
        $S_z^s \sim \mathcal{U}(0.85, 1.15)$
        $\alpha \sim \mathcal{U}(0, \pi)$
        $\mathcal{V}'_{cyl} \leftarrow \{f(\vec{v}; S_x^c, S_y^c, S_z^c, \alpha) | \vec{v} \in \mathcal{V}_{cyl}\}$
        $\mathcal{V}'_{sph} \leftarrow \{f(\vec{v}; 1, 1, S_z^s, 0) | \vec{v} \in \mathcal{V}_{sph}\}$
        $t_x, t_y \sim \mathcal{U}(0, 0.1)$          ▷ draw position offset
        for all $\vec{v} \in \mathcal{V}'_{cyl} \cup \mathcal{V}'_{sph}$ do
            $N_x, N_y, N_z \sim \mathcal{N}(0, 0.04)$          ▷ draw AWGN
            $\vec{v} \leftarrow \vec{v} + (N_x, N_y, N_z)^T$
            $\vec{v} \leftarrow \vec{v} + (l_{i,x}, l_{i,y}, 0)^T$          ▷ move to grid pos. $l_i$
            $\vec{v} \leftarrow \vec{v} + (t_x, t_y, 0)^T$          ▷ add position offset
        end for
        renderer.addMesh($\mathcal{V}'_{cyl} \cup \mathcal{V}'_{sph}, \mathcal{F}$)
    end for
    return renderer.getDepthImages()
end procedure
```

III. Evaluation

We evaluate our approach on the data set introduced in previous work [5]. The data set contains 2200 temporal frames, recorded from three commodity stereo-vision-based depth sensors. Each temporal frame consists of three foreground segmented depth images. For the evaluation of our approach we use the same discrete ground floor grid as proposed in [5], resulting in a grid with 15×12 grid points, corresponding to horizontal and vertical distances of $33\,cm$ cm between adjacent grid points. As input of the proposed CNN architecture we used subsampled depth images with a resolution of 188×120 pixel.

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 true positives, false positives and false negatives, respectively. The F1-Score is defined as the harmonic mean of precision and recall, $F_1 = (2 \times \text{precision} \times \text{recall})/(\text{precision} + \text{recall})$. For the evaluation a detection is assumed to be a true positive if it is within a radius of $30\,cm$ of the ground truth. We compare the precision-recall performance of the following methods:

- **Randomized (ours)** refers to the proposed CNN-architecture trained on randomized synthetic depth images (see Fig. 2(b)-(e)).
- **Static (ours)** refers to the proposed method trained only on synthetic depth images (see Fig. 2(a)) based on the static person model as used by [5].
- **MF-Inference** [5] State-of-the-art probabilistic method which uses the same static generative scene model. The

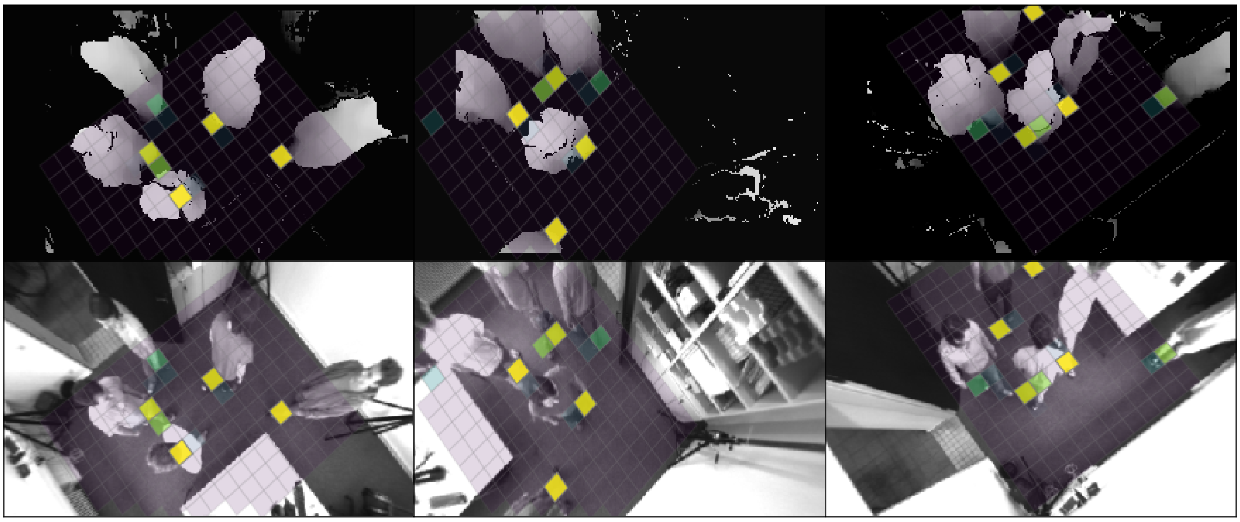

Fig. 3. Exemplary result of our approach (with training data randomization). The output of the proposed neural network is projected onto the ground floor, where purple correspond to a probability of zero and yellow to one respectively.

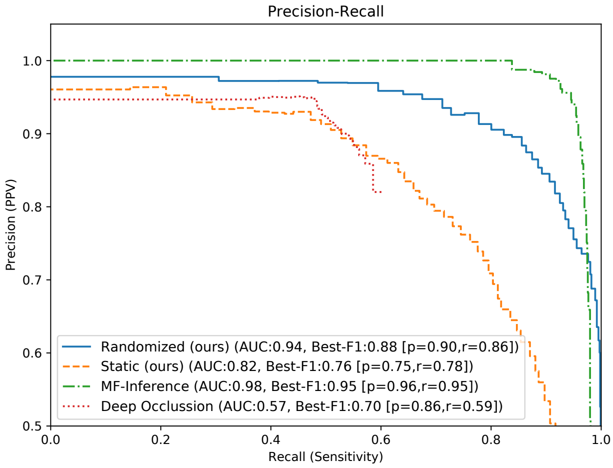

Fig. 4. Precision-Recall curves showing the performance of our approach with and without domain randomization.

output occupancy map layout is identical to the method proposed in this work.

- **Deep Occlusion** [4] is a state-of-the-art end-to-end architecture for monocular multi-view person detection. We use the available pre-trained model. As input, we stack the given gray scale observations to a three channel image to be compatible with the RGB architecture.

Fig. 4 shows the precision-recall performance of the evaluated approaches for all frames. While the computationally intensive stochastic optimization introduced in [5] outperforms all other methods, our proposed method with randomization achieves remarkable results with best F1-score of 0.88. Comparing the two manifestations of the proposed architecture shows that the performance could be significantly increased by randomizing the scene model (F1-score of 0.76 for static person model vs. F1-score of 0.88 for randomized person model). Notice that the given data set is challenging for Deep Occlusion [4] since the method operates on the intensity images and is not trained with any data representing the specific top-view scenario. Fig. 3 shows a typical result of our method.

IV. Conclusion

In this work we have proposed a multi-view CNN architecture to detect people in multiple overlapping depth images from the top-view. In contrast to prevalent methods in the literature our CNN architecture is trained only on synthetic depth images, sampled from a randomized generative scene model. Future work will focus on more realistic scene models as well as the combination of data-driven CNN architectures with state-of-the-art probabilistic models.

References

[1] Q. You and H. Jiang, "Real-time 3d deep multi-camera tracking," *arXiv Prepr. arXiv2003.11753*, Mar. 2020.

[2] Q. Zhang and A. B. Chan, "Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 8289–8298, 2019.

[3] T. Chavdarova and F. Fleuret, "Deep multi-camera people detection," in *Proc. IEEE Int. Conf. Mach. Learn. Appl. ICMLA*, vol. 2017-Decem, pp. 848–853, 2017.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.