Towards Global People Detection and Tracking using Multiple Depth Sensors

Johannes Wetzel, Samuel Zeitvogel, Astrid Laubenheimer
Intelligent Systems Research Group (ISRG)
Karlsruhe University of Applied Sciences
Karlsruhe, Germany
{johannes.wetzel,samuel.zeitvogel,astrid.laubenheimer}@hs-karlsruhe.de

Michael Heizmann
Institute of Industrial Information Technology (IIIT)
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
michael.heizmann@kit.edu

***Abstract*—In this work a novel approach for multi depth sensor person detection and tracking from top view is presented. We propose a probabilistic framework formulating the problem of people detection in multiple overlapping depth images as an inverse problem. As a generative forward model, we employ a simple differentiable 3D person model allowing us to detect people from arbitrary viewpoints. Furthermore, we extend our probabilistic framework to allow for tracking of individuals over time. Finally, we show how to solve for the global person trajectories exploiting differentiable rendering. The preliminary evaluation shows promising qualitative results of our approach on samples of three stereo vision based depth sensors observing an indoor scene.**

***Index Terms*—multi camera person detection and tracking; multi sensor fusion; network of depth cameras; inverse graphics; inverse problem; generative model; differentiable rendering**

I. Introduction

Tracking and detection of people in a network of cameras is a vital preprocessing task for many applications such as video surveillance, sport game analysis, ambient assisted living, etc. The vast majority of the related literature focuses on the classical video surveillance scenario, capturing the pedestrians from profile and frontal view using 2D video cameras. In this work we focus on a different setup. We address the problem of indoor people detection and tracking using a network of stereo vision based depth sensors. In contrast to the classical video surveillance scenario, the sensors capture the scene from a top view to resolve occlusion in crowded scenes. Due to drastic view point change the persons strongly vary in their appearance, making it very challenging for classical discriminative pedestrian detectors. Moreover, the majority of existing literature of multi camera tracking relies on a local detection approach, where the local detections from every sensor are merged into a global coordinate system. Instead, we formulate the problem of people detection and tracking in multiple depth images as inverse problem, seeking to overcome the challenges mentioned above in two ways:

1) Generative person model: In contrast to discriminative pedestrian detection algorithms we employ a generative 3D person model, leading to a view-point independent detector which does not depend on a huge training set.
2) Global detection and tracking: Our approach models the detection and tracking with a global probability distribution conditioned on all observations at every time step. Thus, the given dependency between all camera views can be exploited in a natural way. Our approach is capable of integrating redundant as well as complementary information from different views, attempting to resolve occlusion as well as measurement noise.

II. Related Work

Multi camera people detection and tracking has been widely studied in the context of video surveillance. The vast majority of those approaches are based on multiple 2D video cameras observing an outdoor scene. However the topic of people detection and tracking with a network of depth cameras, especially in top down view, is not well studied yet. Therefore we will first discuss the different methods of the classical video surveillance literature and later focus on the more specialized depth sensor setup we address in this work. The task of tracking people in a multi camera network can be divided into approaches working across non-overlapping views [1] and overlapping views [2]. In this work we will only focus on approaches working across overlapping views. For a comprehensive review of multi camera people tracking we refer to [3], [4]. For the rest of this section we categorize the literature on multi camera person tracking in approaches based on local detection and tracking, homography based approaches and inverse problem approaches.

Since detecting and tracking people in a single camera image has been intensivly studied [5], [6] a lot of approaches rely on fusing local detections or local tracklets into a common coordinate system. The fusion of the tracklets of different cameras can be achieved by spatio temporal features [7] and appearance based features [8]. Since each camera is detecting people independently, those approaches do not make full use of the given multi view information, making it very difficult to handle occlusion. Moreover, the applied standard people detectors are optimized to detect pedestrians in frontal view and profile view but not in the top view [9] making them insufficient for our setup.

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

Homography based approaches project local image cues, e.g. single points of interest, pixel intensities or the silhouette of a blob on a common plane and fuse this representation across all views to get global detections [10]. Eshel et al. [11] project the foreground pixels of all views into common height planes to detect the heads of people. Kahn et al. [2] focuses on the region around the foot point of people proposing the *homographic occupancy constraint* to handle occlusion. Peng et al. [12] extends those approaches by a multi view Bayesian network to avoid ”phantom” detections due to heavy occlusion. However, all mentioned approaches rely heavily on a good 2D foreground segmentation, hence suffering from noise, shadows and illumination changes.

A more generic class of approaches formulates the problem of multi camera people detection as an inverse problem by employing a generative person model and minimize the difference between the image evidence of all cameras and the synthetic images. Fleuret et al. [13] propose a probabilistic framework for multi camera people detection and tracking using a simple generative person model in form of a rectangular bounding box. Alahi et al. [14] present a similar approach using a more complex silhouette as a generative forward model. In contrast to our method, both approaches employ only 2D forward models and fit the model to a binary foreground mask.

Tseng et al. [15] present an indoor surveillance system, based on multiple top view kinect depth cameras. For each camera they obtain a virtual top view depth image based on the given point cloud, finally stitching a global depth image. Moreover, they apply a hemiellipsoidal head model to detected people in the global depth image. However, the approach relies on high quality depth data and is limited to the top view.

The present work is inspired by [13] but in contrast we are using depth images as evidence and propose a differentiable generative 3D model using OpenDR [16], allowing us to effectively detect people in arbitrary viewpoints. To the best of our knowledge, the problem of people detection and tracking in multiple top view depth images has not yet been formulated as inverse problem using a differentiable generative 3D model.

III. Approach

We formulate the people detection and tracking problem as an inverse problem using a generative 3D person model. In the literature this method is also referred to as *vision-as-inverse-graphics* or *analysis-by-synthesis*. We assume that the sensors are intrinsically and extrinsically calibrated in advance. Let $\mathbf{P}_c = \mathbf{K}_c[\mathbf{R}_c|\vec{t_c}]$ be the projection matrix for each camera c, with $\mathbf{K}_c$ being the intrinsic camera matrix and $[\mathbf{R}_c|\vec{t_c}]$ the extrinsic transformation which maps a point from the common world coordinate system to the corresponding camera coordinate system. For convenience, we assume that we know the number of people n in the scene a priori. In a practical system, this limitation can be overcome by applying a 2D detector or make use of an iterative scheme trying to find the true value for n. For better understanding, we will first discuss the detection framework for a single time step and afterwards extend it, incorporating a sequence of frames.

Let n be the number of people in the scene and $\vec{X} = (\vec{x}_1, \ldots, \vec{x}_n)$ be the vector describing the person locations $\vec{x} \in \mathbb{R}^2$ in ground plane world coordinates. Our goal is to infer how probable a scene configuration $\vec{X}$ explains the given observations $\vec{O} = (O_1, \ldots, O_C)$ from C cameras at the same time step. In this work we use depth images as observations. Applying Bayes' theorem we get the posterior distribution

$$p(\vec{X}|\vec{O}) = \frac{p(\vec{O}|\vec{X})p(\vec{X})}{p(\vec{O})}. \qquad (1)$$

A. Likelihood

We assume that, the views are independent for a fixed configuration $\vec{X}$. Thus, the likelihood factorizes as follows:

$$p(O_1, ..., O_C|\vec{X}) = \prod_{c=1}^{C} p(O_c|\vec{X}). \qquad (2)$$

We model the likelihood using a generative forward model $G(\cdot)$ 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 camera c using a simple 3D person model. The model consists of a cylinder for the body and a sphere for the head, see Fig.1(c). For rendering, we use the differentiable renderer OpenDR [16]. Assuming that our given observations suffer from Gaussian noise, we can formulate the likelihood as

$$p(O_c|\vec{X}, \sigma) \propto \exp\left(-\frac{1}{2\sigma^2}||O_c - G(\vec{X}, \mathbf{P}_c)||^2\right). \qquad (3)$$

We incorporate the physical model of the sensor in a natural way in to our framework, allowing us to detect people from arbitrary viewpoints and to easily integrate a new sensor modality into the network.

B. Prior

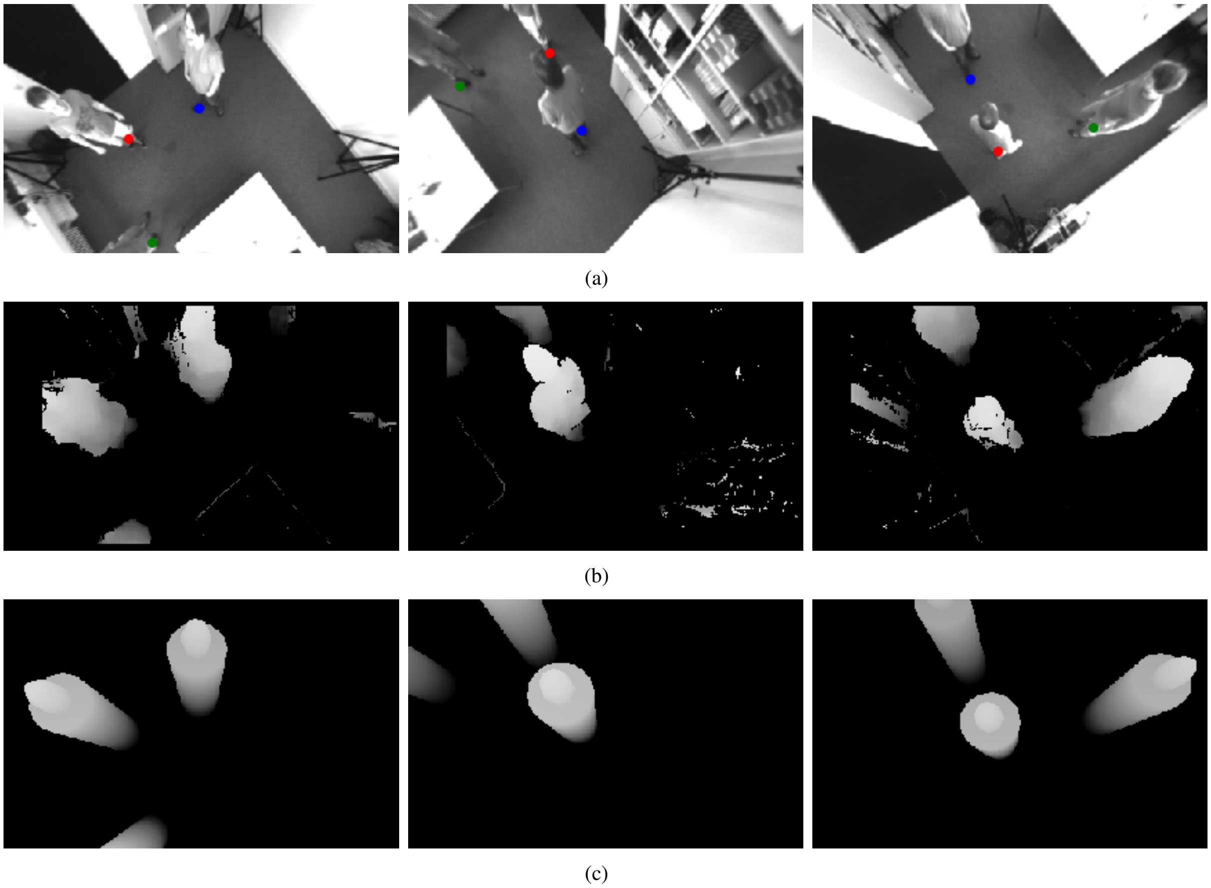
For the detection we employ two independent priors, $p_{\text{box}}(\vec{X})$ and $p_{\text{dist}}(\vec{X})$, hence $p(\vec{X}) = p_{\text{box}}(\vec{X})p_{\text{dist}}(\vec{X})$. Since we can estimate the visible ground plane of the sensor network, we model this knowledge as a uniform distributed prior for every person location in the observable rectangular area, thus we can write

$$p_{\text{box}}(\vec{X}) = \prod_{i=1}^{n} p(\vec{x}_i), \quad p(\vec{x}_i) = \mathcal{U}(\vec{x}_{min}, \vec{x}_{max}) \qquad (4)$$

with $\mathcal{U}(\vec{x}_{min}, \vec{x}_{max})$ being the uniform distribution over the approximated rectangular area given by the interval $[\vec{x}_{min}, \vec{x}_{max}]$. The second prior exploits the assumption that two individuals are keeping a certain distance to each other. Since we assume that the probability $p_{\text{dist}}(\vec{X})$ depends only on the joint probabilities of all possible location pairs, we model the second prior as

$$p_{\text{dist}}(\vec{x}_i, \ldots, \vec{x}_n) = \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} p(\vec{x}_i, \vec{x}_j). \qquad (5)$$

Modelling the joint probability between two locations as a zero mean Gaussian with respect to the inverse distance $d(\vec{x}_i, \vec{x}_j) =$


(a)

(b)

(c)

Fig. 1. Example from an indoor sequence with three sensors. A column corresponds to one sensor, row (a) shows the camera images with reprojected foot points of the inferred detections, (b) shows the observed depth images after background subtraction, used as input for our approach, (c) shows the corresponding synthetic depth images.

$1/(||\vec{x}_i - \vec{x}_j|| + \epsilon)$ we can write the pairwise joint probability distribution as

$$p(\vec{x}_i, \vec{x}_j|\sigma) \propto \exp\left(-\frac{1}{2\sigma^2}||d(\vec{x}_i, \vec{x}_j)||^2\right). \qquad (6)$$

C. Maximum a posteriori estimation

To find the *maximum a posteriori (MAP)* scene configuration as defined in (1) we solve the following non-linear least-squares problem:

$$\begin{aligned}\vec{X}^* &= \arg\max\nolimits_{\vec{X}} p(\vec{X}|O_1, \dots, O_c) \\ &= \arg\min\nolimits_{\vec{X}} \sum_{c=1}^{C} ||O_c - G(\vec{X}, \mathbf{P}_c)||^2 + \alpha E_{\text{box}} + \beta E_{\text{dist}}\end{aligned} \qquad (7)$$

We approximate the box prior in (4) in a way that is computational preferable for continuous numerical optimization. Let $x_{i,r}$ be r-th component of $\vec{x}_i$, then the box penalty is given as

$$\begin{aligned}E_{\text{box}} = \sum_{i=1}^{n} \sum_{r \in \{1,2\}} &[\max(x_{min,r} - x_{i,r}, 0) \\ &+ \max(x_{i,r} - x_{max,r}, 0)]^2.\end{aligned} \qquad (8)$$

The distance energy term is a direct result of the prior proposed in (5), with an additional $\max$ function, which makes sure that the costs are zero for all pairwise distances larger than $\delta = 1m$:

$$E_{\text{dist}} = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left[\max\left(d(\vec{x}_i, \vec{x}_j) - \delta^{-1}, 0\right)\right]^2. \qquad (9)$$

The final non linear least squares problem is solved using a trust region method like dogleg, exploiting the gradients provided by the differentiable renderer framework OpenDR [16]. For better convergence we apply a coarse to fine strategy using a Gaussian image pyramid.

D. Tracking

So far, we have discussed inferring the locations $\vec{X}^t$ of people given the observation of all cameras for one time step t. Given a time series of m observations $\mathcal{O} = (\vec{O}^1, \dots, \vec{O}^m)$ we can extend the framework defined above to infer the location of every individual at every time step $\mathcal{X} = (\vec{X}^1, \dots, \vec{X}^m)$. We rewrite the posterior distribution in (1) as

$$p(\mathcal{X}|\mathcal{O}) = \frac{\left[\prod_t \prod_c p(O_c^t|\vec{X}^t)\right] p(\mathcal{X})}{p(\mathcal{O})}. \qquad (10)$$

We employ an additional prior $p_{\text{time}}(\mathcal{X})$ which represents the dynamics between consecutive observations. Since we assume only small movements between two frames, we use a simple Markov model

$$p_{\text{time}}(\mathcal{X}) = \prod_{t=2}^{m}\prod_{i=1}^{n} p(\vec{x}_i^t|\vec{x}_i^{t-1}) \tag{11}$$

assuming that a person location at time step $\vec{x}^t$ is just a noisy version of its predecessor $\vec{x}^{t-1}$, thus we can write

$$p(\vec{x}_i^t|\vec{x}_i^{t-1}, \mathbf{\Sigma}) = \mathcal{N}(\vec{x}_i^{t-1}, \mathbf{\Sigma}). \tag{12}$$

Since the likelihood and prior terms defined in (2) and (4,5) respectively are staying the same for every single time step, we only need to sum those up over time and add the additional energy term for the prior $p_{\text{time}}(\mathcal{X})$. The total MAP objective for global detection and tracking is then given as

$$\begin{aligned} \mathcal{X}^* = \arg\min\nolimits_{\mathcal{X}} \sum_{t=1}^{m}\sum_{c=1}^{C} ||O_c^t - G(\vec{X}^t, \mathbf{P}_c)||^2 \\ +\alpha\sum_{t=1}^{m} E_{\text{box}} + \beta\sum_{t=1}^{m} E_{\text{dist}} + \gamma\sum_{t=2}^{m} ||\vec{X}^t - \vec{X}^{t-1}||^2. \end{aligned} \tag{13}$$

IV. Evaluation

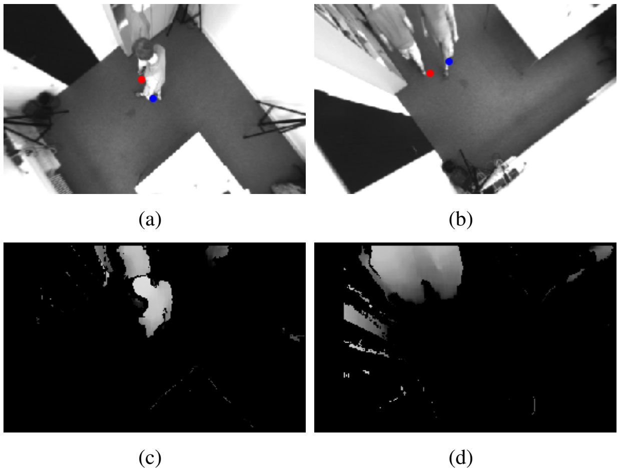
Due to the lack of a publicly available dataset for multi depth sensor people detection and tracking in top view, we present preliminary qualitative experiments, showing the applicability of our approach. We evaluate our approach on an indoor office scene recorded from three stereo vision based depth sensors. The sensors have a top view on the scene and are mounted at a height of three meters, having a significant overlap to each other, see Fig. 1(a). As input observations we use foreground depth images in a resolution of 376 × 240, obtained by static background subtraction, see Fig. 1(b). We use $\alpha = 0.1, \beta = 0.01, \gamma = 0.01$ as weight parameters for the regularization terms in (13).

Fig. 1 illustrates the MAP solution for one sample frame. Our generative model (Fig. 1(c)) is able to explain the given observations (Fig. 1(b)) quite well. Notice how our approach makes use of all image evidence, even if people are only partially visible. This improves the global detection result. In Fig. 2 we use the image evidence of two sensors only. People are correctly located, even when heavy occlusion is present (Fig. 2(a)) and the viewpoint is quite extreme (Fig. 2(b)).

V. Conclusion

In the present work we have addressed the problem of people detection and tracking in multiple overlapping depth images as an inverse problem. We have proposed a natural probabilistic formulation for people detection and tracking, using a generative 3D person model. Moreover, we have shown how to solve for the global MAP solution, exploiting differentiable rendering.

Future work will include a quantitative evaluation of the proposed approach as well as the investigation of approximate bayesian inference methods such as *Markov chain Monte Carlo* or *variational inference* to not only approximate the full posterior distribution but also overcome the shortcomings of the proposed gradient based optimization strategies, e.g. the a priori needed number of persons in the scene.



(a)

(b)

(c)

(d)

Fig. 2. Example with two sensors. A column corresponds to one sensor. Fig. (a) and (b) show the camera images, (c) and (d) the observed depth images.

References

[1] A. Rahimi, B. Dunagan, and T. Darrell, "Simultaneous calibration and tracking with a network of non-overlapping sensors," in *CVPR*, vol. 1, pp. 187–194, 2004.

[2] S. M. Khan and M. Shah, "Tracking multiple occluding people by localizing on multiple scene planes," *TPAMI*, vol. 31, no. 3, 2009.

[3] X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern Recognition Letters*, vol. 34, pp. 3–19, jan 2013.

[4] L. Hou, W. Wan, J.-N. Hwang, R. Muhammad, M. Yang, and K. Han, "Human tracking over camera networks: a review," *EURASIP Journal on Advances in Signal Processing*, no. 1, p. 43, 2017.

[5] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Towards Reaching Human Performance in Pedestrian Detection," *TPAMI*, vol. 8828, no. c, pp. 1–1, 2017.

[6] H. Kieritz, S. Becker, W. Hubner, and M. Arens, "Online multi-person tracking using Integral Channel Features," *AVSS*, pp. 122–130, 2016.

[7] N. Anjum and A. Cavallaro, "Trajectory association and fusion across partially overlapping cameras," *AVSS*, pp. 201–206, 2009.

[8] G. Kayumbi, N. Anjum, and A. Cavallaro, "Global trajectory reconstruction from distributed visual sensors," *ACM/IEEE International Conference on Distributed Smart Cameras, ICDSC*, 2008.

[9] C. Ertler, H. Possegger, M. Opitz, and H. Bischof, "Pedestrian Detection in RGB-D Images from an Elevated Viewpoint," in *22nd Computer Vision Winter Workshop, CVWW*, 2017.

[10] B. A. C. Sankaranarayanan, A. Veeraraghavan, and R. Chellappa, "Object Detection, Tracking and Recognition for Multiple Smart Cameras," *Proceedings of the IEEE*, vol. 96, no. 10, 2008.

[11] R. Eshel and Y. Moses, "Homography based multiple camera detection and tracking of people in a dense crowd," *CVPR*, 2008.

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

[13] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *TPAMI*, vol. 30, 2008.

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

[15] T. E. Tseng, A. S. Liu, P. H. Hsiao, C. M. Huang, and L. C. Fu, "Realtime people detection and tracking for indoor surveillance using multiple top-view depth cameras," *IROS*, pp. 4077–4082, 2014.

[16] M. M. Loper and M. J. Black, "OpenDR: An approximate differentiable renderer," *ECCV*, vol. 8695 LNCS, no. PART 7, pp. 154–169, 2014.