Probabilistic Models and Inference for Multi-View People Detection in Overlapping Depth Images

Zur Erlangung des akademischen Grades eines

DOKTORS DER INGENIEURWISSENSCHAFTEN (Dr.-Ing.)

von der KIT-Fakultät für

Elektrotechnik und Informationstechnik

des Karlsruher Instituts für Technologie (KIT)

angenommene

DISSERTATION

von

Johannes Wetzel, M.Sc.

geb. in Gengenbach

Tag der mündl. Prüfung:	21. Oktober 2021
Hauptreferent:	Prof. Dr.-Ing. Michael Heizmann, KIT
Korreferent:	Prof. Dr.-Ing. Astrid Laubenheimer, HKA

Dedicated to Mia, Lene and Janne

Preface

The present work was accomplished while being a research assistant at the Intelligent System Research Group (ISRG), HKA in cooperation with the Institute of Industrial Information Technologies (IIIT), KIT. While many people accompanied me during my time as a PhD student my gratitude is especially to those mentioned below.

First of all, I want to thank Prof. Dr.-Ing. Michael Heizmann for being a great supervisor, guiding me on my path to the present work. A special thank goes to Prof. Dr.-Ing. Astrid Laubenheimer for not only supervising this thesis but also giving me the opportunity to be part of her research group and supporting my path in academia as a mentor.

Furthermore, I want to thank my fellow PhD students at the IIIT for the nice working atmosphere and their generous support. Within ISRG, I want to thank Prof. Dr. Norbert Link, Martin Redlof, Anita Bender, Alexander Melde and Christian Wernet. A special thanks goes to my fellow PhD students Samuel Zeitvogel, Johannes Dornheim and Tarek Iraki. I'm very grateful for our time together at ISRG, including the countless fruitful discussions but also the fun we had while not at work.

This thesis would have not been possible without my family. A special thanks goes to my brother Micha, who accompanied me during my time as a PhD student. I also want to thank my parents Susanne and Ralf for their generous and unconditional support. My deepest gratitude is due to my wife Rike, for her endless patience and mental support during the intense time of this work.

Karlsruhe, November 2021 Johannes Wetzel

Zusammenfassung

Die sensorübergreifende Personendetektion in einem Netzwerk von 3D-Sensoren ist die Grundlage vieler Anwendungen, wie z.B. Personenzählung, digitale Kundenstromanalyse oder öffentliche Sicherheit. Im Gegensatz zu klassischen Verfahren der Videoüberwachung haben 3D-Sensoren dabei im Allgemeinen eine vertikale top-down Sicht auf die Szene, um das Auftreten von Verdeckungen, wie sie z.B. in einer dicht gedrängten Menschenmenge auftreten, zu reduzieren. Aufgrund der vertikalen top-down Perspektive der Sensoren variiert die äußere Erscheinung von Personen sehr stark in Abhängigkeit von deren Position in der Szene. Des Weiteren sind Personen aufgrund von Verdeckungen, Sensorrauschen sowie dem eingeschränkten Sichtfeld der top-down Sensoren häufig nur partiell in einer einzelnen Ansicht sichtbar.

Um diese Herausforderungen zu bewältigen, wird in dieser Arbeit untersucht, wie die räumlich-zeitlichen Multi-View-Beobachtungen von mehreren 3D-Sensoren mit sich überlappenden Sichtbereichen effektiv genutzt werden können. Der Fokus liegt insbesondere auf der Verbesserung der Detektionsleistung durch die gemeinsame Betrachtung sowohl der redundanten als auch der komplementären Multi-Sensor-Beobachtungen, einschließlich des zeitlichen Kontextes. In der Arbeit wird das Problem der Personendetektion in einer Sequenz sich überlappender Tiefenbilder als inverses Problem formuliert. In diesem Kontext wird ein probabilistisches Modell zur Personendetektion in mehreren Tiefenbildern eingeführt. Das Modell beinhaltet ein generatives Szenenmodell, um Personen aus beliebigen Blickwinkeln zu erkennen. Basierend auf der vorgeschlagenen probabilistischen Modellierung werden mehrere Inferenzmethoden untersucht, unter anderem Gradienten-basierte kontinuierliche Optimierung, *Variational Inference*, sowie *Convolutional Neural Networks*. Dabei liegt der Schwerpunkt der Arbeit auf dem Einsatz von Variationsmethoden wie *Mean-Field Variational Inference*. In Abgrenzung zu klassischen Verfahren der Literatur wird hier keine Punkt-Schätzung vorgenommen, sondern die a-posteriori Wahrscheinlichkeitsverteilung der in der Szene anwesenden Personen approximiert. Durch den Einsatz des generativen Vorwärtsmodells, welches die Charakteristik der zugrundeliegenden Sensormodalität beinhaltet, ist das vorgeschlagene Verfahren weitestgehend unabhängig von der konkreten Sensormodalität.

Die in der Arbeit vorgestellten Methoden werden anhand eines neu eingeführten Datensatzes zur weitflächigen Personendetektion in mehreren sich überlappenden Tiefenbildern evaluiert. Der Datensatz umfasst Bildmaterial von drei passiven Stereo-Sensoren, welche eine top-down Sicht auf eine Bürosituation vorweisen. In der Evaluation konnte nachgewiesen werden, dass die vorgeschlagene Mean-Field Variational Inference Approximation Stand-der-Technik-Resultate erzielt. Während Deep Learnig Verfahren sehr viele annotierte Trainingsdaten benötigen, basiert die in dieser Arbeit vorgeschlagene Methode auf einem expliziten probabilistischen Modell und benötigt keine Trainingsdaten. Ein weiterer Vorteil zu klassischen Verfahren, welche häufig nur eine MAP Punkt-Schätzung vornehmen, besteht in der Approximation der vollständigen Verbund-Wahrscheinlichkeitsverteilung der in der Szene anwesenden Personen.

Abstract

Wide-area indoor people detection in a network of depth sensors is the basis for many applications, e. g. people counting, customer behavior analysis, public security or ambient assisted living. In contrast to classical pedestrian detection approaches, depth sensors typically capture the scene from the top-view to minimize occlusions in crowded scenes. As a consequence of the vertical top-view, position changes of individuals lead to drastically varying appearances. This makes the people detection task quite challenging for off-the-shelf, data-driven pedestrian detectors. Moreover, people are occasionally only partially visible in a single view due to occlusion, measurement noise or the limited field of view of a top-view depth sensor.

Considering the aforementioned challenges, the present thesis examines how to exploit the full temporal multi-view image evidence. In particular, we investigate how the redundant and complementary multi-view information, including the temporal context, can be jointly leveraged to improve the detection performance. We recast the problem of multi-view people detection in a sequence of overlapping depth images as an inverse problem and present a generative probabilistic framework to jointly exploit the temporal multi-view image evidence. Based on the proposed probabilistic model, we examine several inference methods, including continuous gradient based optimization, variational inference and end-to-end deep learning methods. As our main contribution, we propose to use mean-field variational inference to not only estimate the maximum a posteriori (MAP) state, but to also approximate the joint posterior probability distribution of people present in the scene across space and time.

For evaluation, we introduce a novel data set for indoor people detection in multiple overlapping top-view depth images. We report state-of-the art results for the proposed mean-field variational inference methods. Furthermore, we demonstrate that, compared to the frame-by-frame mono-view setup, our approach successfully exploits the temporal multi-view image evidence and robustly converges in only a few iterations.

Contents

Nomenclature

Common Abbreviations

Abbreviation	Description
AWGN	Additive white Gaussian noise
CAVI	Coordinate ascent variational inference
cf.	Short form of Latin confer, meaning compare with
CNN	Convolutional neural network
CRF	Conditional random fields
DNN	Deep neural network
ELBO	Evidence lower bound
HOG	Histogram of oriented gradients
i.i.d.	Independently identically distributed
KL	Kullback-Leibler (divergence)
MAP	Maximum a posteriori
MF-VI	Mean-field variational inference
ML	Maximum likelihood
NLLSQ	Non-linear least squares
PDF	Probability density function
PMF	Probability mass function
POM	Probabilistic occupancy map
SVM	Support vector machine
VI	Variational inference

Letters

Latin Letters

Symbol	Description
$\bar{A}_c^{i=s}$	Conditional synthetic average image in perspective of camera c with respect to the current mean-field state and x_i forced to state $s \in \{0, 1\}$ (POM)
$\bar{A}_c$	Synthetic average image in perspective of camera c with respect to the current mean-field state (POM)
$\mathcal{A}_c^i$	Synthetic binary image with a rectangle placed at location i in perspective of camera c (POM)
b_c	Foreground segmented binary image from sensor c (POM)
$\mathbf{b}$	Vector of binary foreground images $\mathbf{b} = (b_1, \ldots, b_C)^\intercal$ used as observations (POM)
C	Number of sensors
E_t^{box}	Energy term corresponding to the box prior in MAP inference at time step t
E_t^{dist}	Energy term corresponding to the distance prior in MAP inference at time step t
$\mathcal{F}$	Set of faces of a 3D person model
g	Synthetic depth image
H	Image height in pixels
h	Realization of a uniform random variable, reflecting the number of expected persons in a randomly drawn scene configuration
$I_c[u_i]$	Rectangular bounding box of the 3D model at location u_i in perspective of sensor c
J	Number of threshold values ρ used for precision-recall evaluation
l_{bce}	Binary cross-entropy loss for multi-view CNN architecture
$l_{i,x}, l_{i,y}$	Ground plane x- and y-coordinates of discrete grid location u_i
m	Number of people present in the scene (continuous latent space)
n	Number of discrete ground plane grid cells (also referred to as grid locations) u_i
N_i'	Set of the direct neighbor indices of a grid cell with index i, including the index i

Symbol	Description
N_i	Set of the direct neighbor indices of a grid cell with index i, excluding the index i
$o_{c,t}$	Foreground segmented depth observation from sensor c and time t
$\mathcal{O}$	Symbolic representation for an observation
$\mathbf{o}_t$	Vector of foreground segmented depth observations $\mathbf{o}_t = (o_{1,t}, \ldots, o_{C,t})^\intercal$ at time step t
$p_{\text{box}}(\mathfrak{X})$	Box prior distribution corresponding to the continuous probabilistic model
p_{FC}	Dropout retention probability after a fully connected layer block
p_{CNN}	Dropout retention probability after a CNN block
$p_{\text{dist}}(\mathfrak{X})$	Person distance prior distribution corresponding to the continuous probabilistic model
$\mathcal{Q}$	Space of proxy distributions $q \in \mathcal{Q}$ (also referred to as q-family)
$q_i(\cdot)$	Marginal probability distribution for a single latent variable with index i, used as proxy distribution for variational inference
q_i^{init}	Initial mean-field optimization probability for a marginal distribution q_i
$q_{i,t}(\cdot)$	Marginal probability distribution of a person present at location u_i at time step t, used as proxy distribution for variational inference
r	Number of latent variables
$\mathcal{S}$	Symbolic representation of a depth sensor
S_x^c, S_y^c, S_z^c	Uniform random variables representing the scaling of the cylinder component in x-,y- and z-directions
s_x, s_y, s_z	Axis-dependent (x-,y- and z-axis) scaling parameters for parameterized transformation used for randomized scene model
S_z^s	Uniform random variable representing the z-scaling of the sphere component
T	Number of temporal frames in the sequence $1 \ldots T$
t_x, t_y	Realization of uniform random variables, reflecting the position offset in x- and y- direction
u_i	Symbolic representation of a grid location at index i
$\mathbf{v}$	Vertex $\mathbf{v} = (x, y, z)^\intercal$ for definition of a 3D mesh
$\mathcal{V}$	Set of vertices of a single person model $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$

Symbol	**Description**
$\mathcal{V}_{\text{cyl}}$	Set of vertices belonging to the cylinder component of a 3D person model
$\mathcal{V}_{\text{sph}}$	Set of vertices belonging to the sphere component of a 3D person model
w_k	Normalization weight coefficient
W	Image width in pixels
$\mathbf{x}$	Discrete scene configuration $\mathbf{x} \in \{0, 1\}^n$
$\check{\mathbf{x}}$	Location of a person on the ground plane $\check{\mathbf{x}} \in \mathbb{R}^2$
$\check{\mathbf{x}}_{\text{max}}$	Maximum x and y ground plane coordinates (bottom-right corner) corresponding to the box prior with $\check{\mathbf{x}}_{\text{max}} \in \mathbb{R}^2$
$\check{\mathbf{x}}_{\text{min}}$	Minimum x and y ground plane coordinates (top-left corner) corresponding to the box prior with $\check{\mathbf{x}}_{\text{min}} \in \mathbb{R}^2$
$\mathcal{X}$	Symbolic representation of arbitrary latent variables or an abstract scene configuration
$\mathfrak{X}$	Continuous scene configuration with $\mathfrak{X} = (\check{\mathbf{x}}_1, \ldots, \check{\mathbf{x}}_k)$
$x_{i,t}$	Realization of a Bernoulli random variable assigned to grid cell i at time t
x'	One latent variable $x' \in \mathcal{X}$
$\tilde{\mathbf{x}}_i$	Reduced neighborhood scene configuration $\mathbf{x} \in \{0, 1\}^8$
$X_{i,t}$	Bernoulli random variable assigned to grid cell i at time t
$\mathbf{y}$	Discrete ground truth scene configuration $\mathbf{y} = (y_1, \ldots, y_n)^\intercal \in \{0, 1\}^n$ for end-to-end CNN training
$\mathbf{Z}$	Hidden latent random variable
$\mathbf{z}$	One-hot-encoded realization of a random variable $\mathbf{Z}$ with $\mathbf{z} \in (z_1, \ldots, z_n)^\intercal$

Greek Letters

Symbol	**Description**
α	Weight coefficient for asymmetric image similarity
Ω	Random variable representing an abstract scene configuration
ρ	Detection threshold used to generate precision-recall curves

Symbol	Description
η	Random variable representing the measurement noise of the depth observations
κ	Inter person distance threshold for MAP optimization
λ_{box}	Weighting parameter for MAP box energy term E_t^{box}
λ_{dist}	Weighting parameter for MAP distance energy term E_t^{dist}
λ_{future}	Weighting parameter for future expectation $\Psi_{i,t}^{\text{future}}$
λ_{past}	Weighting parameter for past expectation $\Psi_{i,t}^{\text{past}}$
$\lambda_{\text{temporal}}$	Weighting parameter for MAP temporal term
μ	Mean parameter of a probability distribution
ν	Arbitrary local scope variable
ξ	POM parameter reflecting the reliability of the measurement
$\Psi_{c,i}$	Expectation corresponding to the data term for one camera c and grid location i
$\tilde{\Psi}_{c,i}$	Approximated expectation corresponding to the data term for one camera c and grid location i
Ψ_i^{data}	Expectation corresponding to the data term at grid location i
Ψ_i^{filter}	Expectation corresponding to the filtering term at grid location i
$\Psi_{i,t}^{\text{future}}$	Expectation corresponding to the future term at time step t and grid location i
$\Psi_{i,t}^{\text{past}}$	Expectation corresponding to the past term at time step t and grid location i
$\Psi_{c,i}^{\text{pom}}$	Expectation corresponding to the POM data term for camera c and marginal distribution q_i
Ψ_i^{pred}	Expectation corresponding to the predictive term at grid location i
$\tilde{\Psi}_i^{\text{pred}}$	Approximated predictive expectation at grid location i
$\Psi_{i,t}^{\text{smooth}}$	Expectation corresponding to the temporal smoothing term at time step t and grid location i
$\Psi_{i,t}^{\text{temp}}$	Expectation corresponding to the temporal term at time step t and grid location i
Ψ_i	Mean-field update expectation for a single time step corresponding to grid location i
Σ	Covariance matrix of dynamics noise in the continuous latent space
σ	Std. deviation of a normal distribution
σ_{obs}	Std. deviation of depth measurement noise

Symbol	Description
σ_{dist}	Std. deviation of the distance between two individuals
τ_i	Prior term in the mean-field update equation assigned to grid cell i
$\zeta_x, \zeta_y, \zeta_z$	Realization of a normal random variable, reflecting AWGN noise in x, y, z direction respectively
γ	Rotation angle around z-axis for randomization of the generative scene model

Indices

Index	Description
$(\bullet)_{1:T}$	Time sequence from 1 to T
$(\bullet)_c$	Sensor
$(\bullet)_i$	Discrete grid cell on ground floor
$(\bullet)_t$	Time

Mathematical Operators

Expression	Description
$f(\mathbf{v}; \cdot)$	Transformation of a vertex **v** (scaling and rotation around z-axis)
$G_c(\cdot)$	Generative scene model, maps a scene configuration to a synthetic depth image in perspective of sensor c
$A_c(\cdot)$	Generative model (POM), maps a scene configuration to a synthetic binary foreground image in perspective of camera c
$\odot$	Hadarmard product (elementwise multiplication)
$\delta(\cdot, \cdot)$	Image similarity function
$\delta_{\text{asym}}(\cdot, \cdot)$	Asymmetric image similarity function
$\delta_{\text{pom}}(\cdot, \cdot)$	Image similarity function used by probabilistic occupancy map (POM)
$\Vert\cdot\Vert_1$	L1-norm
$\Vert\cdot\Vert_2$	L2-norm
$\text{precision}(\cdot)$	Provides the precision value for a given detection threshold

Expression	Description
$\mathrm{recall}(\cdot)$	Provides the recall value for a given detection threshold
$\mathbf{R}_z(\gamma)$	Rotation matrix, performs a rotation of angle γ around the z-axis
$M(\cdot)$	Threshold function $M : \mathbb{R}^{W\times H} \to \{0,1\}^{W\times H}$ which maps an image to its binary foreground mask
$\overline{M}(\cdot)$	Inverse threshold function

Probabilistic Notation

Expression	Description
$\mathcal{B}(\mu)$	Bernoulli distribution with mean parameter μ
$\langle\cdot\rangle_{p(\cdot)}$	Expected value with respect to the PDF or PMF $p(\cdot)$
$A \perp\!\!\!\perp B$	Denotes that two random variables A, B are statistically independent
$\mathrm{KL}(\cdot \parallel \cdot)$	Kullback–Leibler divergence
$\mathcal{N}\left(\mu, \sigma^2\right)$	Normal distribution with mean parameter μ and variance σ^2
Z	Partition function of an arbitrary PDF or PMF
$p(\cdot)$	Arbitrary PDF or PMF
$q(\cdot)$	Proxy PDF or PMF for variational inference approximation
$\hat{q}(\cdot)$	Optimal proxy PDF or PMF with respect to the given variational inference objective
$\mathcal{U}(a, b)$	Uniform distribution in the interval $[a, b]$

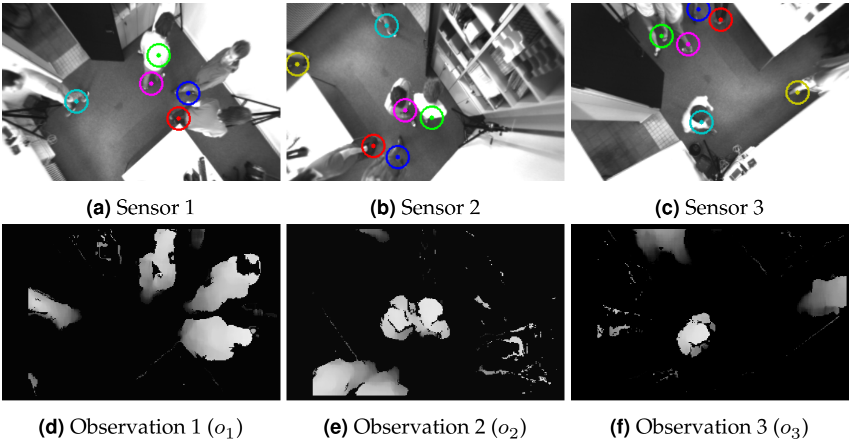
List of Figures

1 Introduction

By virtue of the emergence of low-cost commodity depth sensors[1], there is an increasing demand for privacy-preserving wide-area people detection in various indoor scenarios, e. g. people counting, customer behavior analysis, public security, ambient assisted living or smart homes. In contrast to the classical outdoor video surveillance scenario, the mounting height is very limited in many indoor scenarios. Therefore, the depth sensors typically capture the scene from the top-view to reduce occlusions in crowded scenes. However, as a direct consequence of the top-view and the limited mounting height, the resulting field of view and observable area of a single depth sensor is quite limited. This is an issue in many real-world applications such as customer behavior analysis in shopping malls or airports. Providing complete detection in a wide-area scenario is achieved by employing a network of multiple sensors, covering a larger area.

Apart from the increased observable area, there are additional advantages compared to the classical single-view approach. Since a single view 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. 1.1). Hence, the detection performance (including the detection confidence) in the overlapping regions can be improved by complementary image evidence from multiple views. In particular, this is

[1] A depth sensor refers to a range imaging sensor, which measures the distance to points in the scene.



(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 1.1 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 person marked with green, while in (b) the pair can be clearly separated. In the views (b,c) several people are only partially visible. Note that for inference only the depth images (d-f) are used.

relevant in 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 the computer vision literature. 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 contrast, in this thesis we[2] focus on the task of top-view people detection in a network of low-cost commodity depth sensors.

[2] Unambiguously, the present work is a contribution from the sole author. However, for the ease of an active voice writing style, in this thesis the first-person plural is used instead of the first-person singular. Depending on the context, "we" refers to the sole author or the reader and the author, respectively.

1.1 Problem Statement

In this thesis we focus on indoor people detection in a sequence of consecutive multi-view frames, where each multi-view frame is given by multiple overlapping depth images. The depth images are obtained by a network of low-resolution commodity depth sensors. The sensors have a top-view on the scene with fields of view having a significant joint overlap. Fig. 1.1 shows one exemplary multi-view frame obtained by our setup. The grayscale images in the top-row (Fig. 1.1 (a-c)) only serve for the purpose of visualization. For inference, we use the foreground-segmented depth observations in Fig. 1.1 (d-f). The major challenges are summarized as follows:

- Due to the passive[3] low-resolution stereo sensors, the depth observations suffer from heavy measurement noise, including areas with no measurements.
- As a consequence of the vertical top-view, position changes of individuals lead to drastically varying appearances (see Fig. 1.1). This makes the single view detection task challenging for discriminative detection algorithms, especially for off-the-shelf data-driven pedestrian detectors without a domain-specific large scale data set.
- Frequently, people are only partially visible in one camera view due to occlusion, the limited field of view or measurement noise. However, a typical detector operates independently on a single view and thereby does not make use of the given complementary image evidence. The challenge is to leverage the full multi-view image evidence, including multiple overlapping regions, in order to gain more information about the true state of the scene.
- The depth sensors provide a pseudo synchronized sequence of consecutive frames with approximately 15 frames per second. It is

[3] In contrast to active depth sensors (e. g. time-of-flight or structured light based devices), passive stereo sensors do not emit any signal.

desirable to additionally make use of this temporal information in order to improve the overall detection performance.

To overcome these challenges, the overarching idea followed in this thesis is to jointly make use of the temporal multi-view image evidence. We aim for methods which are able to make use of the given redundant and complementary image evidence, including the temporal context. At the same time we want to minimize the loss of information typically occurring in a cascade of independently applied algorithms, each operating on lossy representations of the output of its predecessor (e. g. a tracking-by-detection pipeline applied to each single view independently). In terms of probability theory, we are interested in methods which approximate the joint probability distribution of people present in a scene across space and time.

1.2 System Overview

To address the challenges identified in Sect. 1.1 we introduce a novel probabilistic framework to approximate the distribution of people present in the scene across space and time. The proposed approach is schematically illustrated in Fig. 1.2. In contrast to the majority of existing multi-view people detection approaches in the literature, we do not rely on locally applied discriminative pedestrian detectors requiring a large scale labeled data set. Instead, we formulate the problem of people detection in multiple overlapping depth images as an inverse problem. Therefore, we introduce a generative scene model, which maps a scene state (in the following referred to as scene configuration) to a synthetic depth image in the perspective of each sensor (see Fig. 1.2 top center). The generative approach allows to effectively handle the different appearances[4] of people, arising from (i) the change of viewpoint and (ii) the partial visibility of

[4] In the computer vision literature *appearance* typically refers to the texture and color of an object. In contrast, in this context it refers to the appearance of an individual in a depth image.

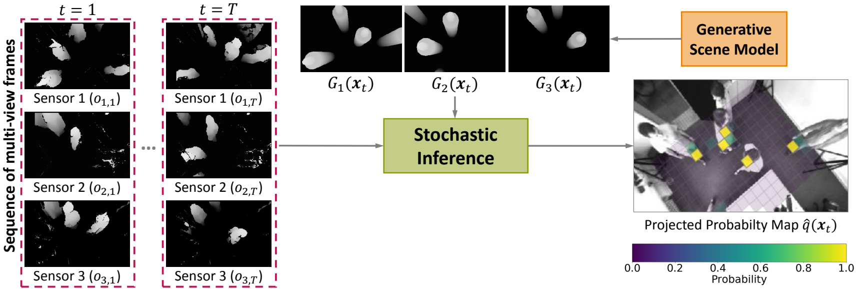


Figure 1.2 Overview of the proposed approach. A time sequence of foreground segmented multi-view depth images from three sensors is used as input (left). The generative scene model generates synthetic depth images with respect to the given intrinsic and extrinsic sensor parameters (middle). The output of the stochastic inference are discrete probability maps representing the probability of people present on the ground level plane across time and space (right).

people, e. g. due to occlusion or the limited field of view. This yields a viewpoint independent detector without the need of a training data set. 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.

Since in general we have access to a sequence of consecutive temporal frames, we extend this framework to additionally consider the temporal context. A common way to leverage temporal information is to use the detections obtained from an isolated multi-view 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 into account the temporal image evidence from all sensor-views. Therefore, we introduce a probabilistic dynamics model to express the probability flow over time. In consequence, we are able to define the full joint distribution of people present in the scene across all sensor views and time steps.

For inference, we propose two different methods. First, we present a maximum a posteriori (MAP) method for people detection. Inference is obtained by a continuous optimization method, which hinges on a good initialization. The MAP approach reveals the shortcomings of point estimates in comparison to methods that explore the full posterior distribution of people present in the scene. Second, as our main contribution we present a variational approach. Discretization of the ground level plane (see Fig. 1.2 right) enables an effective estimate of the posterior distribution. Instead of just estimating an MAP point estimate, we apply mean-field variational inference to approximate the desired joint probability distribution of people present in the scene.

1.3 Contribution

The idea of generative probabilistic modeling has been successfully applied to the task of multi-view people detection with monocular video cameras by Fleuret *et al.* [38]. However, to the best of our knowledge, probabilistic modeling in combination with a generative scene model has not yet been studied in the context of people detection in overlapping depth images. In particular, the scientific contribution of the present thesis can be summarized as follows:

- We introduce a probabilistic model for multi-view people detection in overlapping depth images, including a generative scene model based on an efficient 3D person shape model.

- For continuous inference we deduce the MAP objective and show how to practically solve the resulting non-linear least squares optimization problem by leveraging approximate differentiable rendering [68].

- Our main contribution introduces mean-field variational inference to approximate the posterior distribution of people present in the scene. In contrast to Fleuret *et al.* [38] we use depth images as evidence and therefore are able to employ a richer generative scene model. We also propose a novel strategy to approximate the final mean-field update expectation by making use of geometric scene knowledge and a pre-computed visual dictionary.

- We propose a novel extension to incorporate the temporal context into the mean-field optimization. To that end, we (i) present a probabilistic grid-based dynamics model to define the joint distribution across space and time; and (ii) deduce the mean-field variational inference update equations to efficiently approximate the desired probability distributions.

- For comparison with our probabilistic inference methods, we introduce an end-to-end multi-view CNN architecture. The CNN architecture is only trained with synthetic depth images due to fair comparison with the proposed generative methods.

- To the best of our knowledge, no publicly available data set covering the scenario of top-view people detection in a depth sensor network currently exists. We introduce a novel data set for indoor people detection in multiple overlapping top-view depth images.

- For evaluation, we compare our approach with state-of-the-art monocular multi-view people detection methods. We report state-of-the-art results for the proposed mean-field variational inference methods on the aforementioned data set. Furthermore, we demonstrate that our approach (compared to the mono-view setup) successfully exploits the multi-view image evidence and robustly converges in only a few iterations.

1.4 Thesis Outline

The present thesis is structured as follows: Chapter 2 provides an overview of the most relevant literature related to the general task of multi-view people detection. In Chapter 3 two essential fundamentals for the present work are explained. First, the method of mean-field variational inference (Sect. 3.1) is elaborated. Building on that, the probabilistic occupancy map, which is methodically strongly related to the present work, is explained in detail (Sect. 3.2). In Chapter 4 the probabilistic model, including a general part (Sect. 4.2), as well as manifestations for continuous latent space (Sect. 4.3) and discrete latent space (Sect. 4.4), is introduced. Subsequently, in Chapter 5 different inference methods regarding the proposed probabilistic model are discussed. As introduction, the MAP objective for the continuous latent space is deduced (Sect. 5.1). Practical aspects regarding the resulting continuous non-linear least squares optimization problem are discussed in Sect. 5.1.2. In Sect. 5.2 the main contribution of this thesis is presented. The mean-field update equations are deduced and it is shown how they can be approximated in real-world applications. Complementary to the probabilistic inference methods an end-to-end CNN inference method is presented in Sect. 5.3. In Chapter 6 the results of our experiments are discussed. This includes the introduction of a novel data set (Sect. 6.1), comparative evaluation of the proposed probabilistic people detection approach in discrete latent space (Sect. 6.4), as well as qualitative results for MAP inference in continuous latent space (Sect. 6.6). Finally, Chapter 7 concludes the thesis and provides an outlook on future research directions.

2 Related Work

Multi-camera people detection has been extensively 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 from the top-view, has not yet been explored in detail. Therefore, we discuss the related task of people detection in multiple monocular video cameras. Since many multi-view methods accomplish detection and tracking by fusing local detections or local tracklets into a common world coordinate system, we will also briefly focus on related single-view people detection approaches in Sect. 2.1.

Besides the major categories (single and multi-view), we categorize the literature in methods utilizing monocular video cameras (RGB-based approaches, or intensity based approaches) and depth sensing cameras (depth-based approaches). Since many depth sensors also provide RGB information, many hybrid approaches exist (RGB-D), which will be categorized under the depth-based approaches.

Following the aforementioned categorization of RGB-based and depth-based approaches, we will discuss multi-view approaches in Sect. 2.2. In Sect. 2.2.1 we comprehensively discuss methods focusing on pedestrian detection with multiple monocular cameras, and their relation to our approach. In Sect. 2.2.2 we focus on multi-view people detection methods based on depth information. Finally, we reveal the contribution in relation to the existing literature in Sect. 2.3.

The characteristics of the approaches examined in this chapter are summarized in Tab. 2.1. The binary attribute *top-view* indicates if the corresponding method is evaluated on top-view images in the original publication. The attribute *generative* indicates if some kind of generative modeling is used for the detection component. For the multi-view approaches the additional binary attribute *joint detection* indicates that the multi-view detections are obtained jointly, rather than performing detection on each view independently. This chapter is an extension of previously published work [108, Ch. 2] by the author.

2.1 Single-View People Detection

The problem of pedestrian detection and tracking from the frontal or profile view has been examined in detail in the literature. In this work, however, we focus in particular on sensors having a top-view (also referred to as overhead view in the literature) on the scene. As set forth in Sect. 1.1, people detection from the vertical top-view implies new challenges (e. g. the drastically varying appearance of individuals depending on the view-point) compared to the frontal and profile view. While the detection from the top-view becomes increasingly important in the computer vision community, it still remains a niche research area. In the following we will therefore focus in particular, but not exclusively, on single-view people detection from the top-view. For a comprehensive survey of people detection methods in particular specialized on top-view approaches we refer to [3].

2.1.1 RGB-Based Approaches

The problem of people detection and tracking in a single RGB image has been extensively investigated [54, 62, 66, 73, 104]. However, in this section we will in particular focus on the rather rare approaches which are addressing the problem of people detection from the top-view in a single RGB image.

Table 2.1 Characterization of related single and multi-view people detection approaches. The column *top-view* indicates if a method is in particular evaluated with views from the vertical top-view in the original paper. Parentheses indicate that an attribute cannot be assigned unambiguously.

	Modality	Top -view	Gener -ative	Joint detection	Method keyword
Single-View					
Ahmed *et al.* [4]	RGB	✓	X	-	HOG + SVM
Ahmad *et al.* [1]	RGB	✓	X	-	CNN
Ahmad *et al.* [2]	RGB	✓	X	-	CNN
Ahmed *et al.* [5]	RGB	✓	X	-	CNN
Bagautdinov *et al.* [8]	Depth	X	✓	-	Probab. (MF-VI)
Hacinecipoglu *et al.* [45]	Depth	(✓)	X	-	VHF + SVM
Liu *et al.* [67]	Depth	X	X	-	CNN
Tian *et al.* [95]	Depth	X	X	-	CNN
Li *et al.* [61]	RGB-D	X	X	-	CNN
Rauter [80]	Depth	✓	X	-	SVM
Ertler *et al.* [33]	RGB-D	✓	X	-	CNN
Fuentes-Jimenez *et al.* [41]	Depth	✓	X	-	CNN
Sun *et al.* [92]	RGB-D	✓	X	-	Blob detection
Carletti *et al.* [19]	Depth	✓	X	-	Blob detection
Multi-View					
Xu *et al.* [101]	RGB	X	X	✓	Probabilistic
Sankaranarayanan *et al.* [88]	RGB	X	X	X	Homography
Khan *et al.* [53]	RGB	X	X	✓	Homography
Eshel *et al.* [34]	RGB	(✓)	X	✓	Homography
Peng *et al.* [75]	RGB	X	X	(✓)	Probabilistic
Fleuret *et al.* [39, 38]	RGB	X	✓	✓	Probab. (MF-VI)
Alahi *et al.* [6]	RGB	X	✓	✓	Sparse opt.
Baqué *et al.* [10]	RGB	X	(✓)	✓	CNN + CRF
Chavdarova *et al.* [23]	RGB	X	(✓)	✓	CNN
Hou *et al.* [47]	RGB	X	X	✓	CNN
Tang *et al.* [93]	RGB	X	X	X	CNN 3D pose
Chen *et al.* [24]	RGB	X	X	X	CNN
You *et al.* [102]	RGB	X	X	✓	CNN
Ge *et al.* [42]	RGB	X	✓	✓	Probab. (MCMC)
Zhang *et al.* [103]	RGB	X	X	✓	CNN
Castellano *et al.* [21]	RGB	(✓)	X	✓	CNN
Saputra *et al.* [89]	RGB-D	X	X	X	Kinect skeleton
Sun *et al.* [91]	RGB-D	X	X	X	Kinect skeleton
Munaro *et al.* [69]	RGB-D	X	X	X	HOG
Carraro *et al.* [20]	RGB-D	X	X	X	CNN 3D pose
Tseng *et al.* [96]	Depth	✓	X	✓	3D image stitching

Ahmed *et al.* [4] propose a geometric normalization with respect to the optical center in the image to standardize the appearance of individuals. For detection the authors follow the fundamental ideas of [27] and train a Support Vector Machine (SVM) with histogram of oriented gradients (HOG) features.

More recent methods incorporate deep learning frameworks to accomplish people detection from the top-view. In [1] the pre-trained Single Shot Multi Box Detector [65] is applied to top-view people detection in a single RGB image. Ahmad *et al.* [2] follow a similar idea and apply the pre-trained YOLOv3 Convolutional Neural Network (CNN) [81, 82] to the task of top-view people detection. In [5] these ideas are further developed. The authors introduce a top-view people detection and tracking method based on recent deep learning architectures. For detection the YOLOv3 architecture is re-trained with top-view images. The tracking is based on Simple Online and Realtime Tracking (SORT) [100].

2.1.2 Depth-based Approaches

We will first discuss depth-based methods for people detection in the general frontal or profile view. Bagautdinov *et al.* [8] introduce DPOM, a probabilistic occupancy map for occluded single depth images (cf. Fleuret *et al.* [38]). The authors propose a probabilistic generative model to define the expected distribution of depth values with respect to the presence of an individual. For inference a mean-field variational inference strategy is proposed. Hacinecipoglu *et al.* [45] introduce a pose invariant people detection approach, based on point cloud data. The point cloud is first clustered and head candidates are extracted. For classification an SVM with a Viewpoint Feature Histogram (VHF) [86] feature descriptor is trained. Recently CNN architectures [61, 67, 95] are successfully applied to single view depth image people detection, leveraging data sets with many labeled images for training. However, those mentioned approaches are trained with data sets mostly containing people from the classical frontal or profile view but not from the top-view.

As for monocular RGB approaches, another class of methods focuses on top-view people detection. Rauter *et al.* [80] introduce a novel feature descriptor based on local depth differences for top-view people detection. The feature descriptor is used to train an SVM for head-shoulder detection. To reduce the computational complexity, the resulting detector is only applied to potential head candidates, determined by a local maxima search in the depth image. More recent work by Ertler *et al.* [33] fuses depth and RGB data by combining two CNN streams in a Faster R-CNN [83] architecture. The authors propose a late and mid-layer fusion and report comparative evaluation results for the different fusion methods. Recently, Fuentes-Jimenez *et al.* [41] introduce the CNN architecture DPDnet. The proposed encoder-decoder architecture takes a depth image as input and predicts a confidence map (in image coordinates), where each detection is represented by a 2D normal distribution.

The related problem of people counting with a single depth camera from the top-view has been studied in great detail [19, 30, 92, 105]. 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 implicitly and in a rather small area.

For a comprehensive survey of people tracking in a single RGB-D view (including, but not limited to the top-view) we refer to Camplani *et al.* [17]. Complementary, we refer to Liciotti *et al.* [63] for a narrower review of top-view people detection and counting with RGB-D sensors.

2.2 Multi-View People Detection

2.2.1 RGB-Based Approaches

In this section we focus on people detection with multiple monocular cameras. In order to restrict our scope, we do not consider methods working across non-overlapping views [13, 79, 85] but rather focus on methods utilizing overlapping views. For an exhaustive survey of multi-camera people detection and tracking, we refer to [46, 48, 98].

Since people detection and tracking in single-camera views have been intensively studied [66, 73, 104], many methods accomplish multi-view detection and tracking by fusing local detections or local tracklets into a common world coordinate system [7, 52, 101]. Since the detection is performed independently for each view, those methods do not take full advantage of the multi-view information and thus are often not able to resolve occlusion and measurement noise. Besides, the vast majority of employed pedestrian detectors is optimized to detect people in frontal or profile view (cf. Sect. 2.1).

Homography based approaches project local image features from each sensor into a common plane to perform global detection [88]. In [53] a *homographic occupancy constraint* is proposed to handle occlusion and detect people on a common scene plane. Eshel *et al.* [34] propose a similar approach, projecting the foreground pixels of all views into a common height plane for head detection. In [75] those approaches are extended by a multi-view Bayesian network in order to avoid false positive detections arising from occlusion artifacts.

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.* [39] introduce the idea of estimating the probability of people present on the ground level plane by a binary generative scene model. The authors extend their approach and publish it together with an integrated tracking algorithm in [38]. Since then, the method is referred to as Probabilistic Occupancy Map (POM) in the computer vision literature. POM takes binary foreground images as input and compares them with synthesized binary masks. The binary masks are obtained by a generative model, which represents individuals by a simple rectangular box. The final occupancy map is estimated by mean-field variational inference. For the last decade, POM served as a standard method for multi-view people detection and has been evaluated on many standard benchmarks like PETS [12]. Since the method proposed in this thesis is heavily inspired by POM, it will be discussed in detail in Sect. 3.2. Alahi *et al.* [6] also propose a generative method for people detection in a camera network. They recast the task as a linear inverse problem, regularized by a sparsity constraint on the occupancy grid. Other than in [38], a silhouette person model is proposed. Unlike our approach, both methods utilize only 2D models and fit them to a binary foreground mask.

Due to the massive availability of labeled RGB training data, more recent methods employ CNN architectures for multi-view people detection. Baque *et al.* [10] introduce an end-to-end multi-view people detection architecture. They propose to use the generative model from POM [38] in combination with a discriminatively trained CNN. In order to resolve ambiguities arising from occlusion, the interaction between grid cells is taken into account by Conditional Random Fields (CRFs). The final CNN/CRF architecture can be trained end-to-end by exploiting back mean-field techniques [106]. Chavdarova *et al.* [23] present a CNN architecture enabling end-to-end multi-view probabilistic occupancy map estimation. To overcome the lack of an appropriate multi-view data set, an existing monocular pedestrian data set [32] is used.

In contrast to [10, 23], Hou *et al.* [47] introduce an anchor-free end-to-end multi-view CNN architecture without using generative models or CRFs. The multi-view image evidence is fed into a CNN with shared weights among the views. Intrinsic and extrinsic camera calibrations are used to explicitly project each feature map to the ground plane. The projected feature maps are further aggregated by convolutions, finally predicting the occupancy map on the ground plane. The authors report state-of-the-art results on the challenging WILDTRACK data set [22].

Other research groups focus on integrated tracking and detection systems, leveraging deep learning methods for detection and tracking. Tang *et al.* [93] propose joint multi-view people tracking and pose estimation.

They apply a classical tracking-by-detection scheme on each single view. To associate tracklets across views, appearance and semantic features are used. In a second step single-view human 3D pose estimation is applied based on OpenPose [18]. Finally, the resulting 3D poses are fed back to the multi-view tracker in order to improve the tracklet association across multiple views. Chen *et al.* [24] propose a real-time people tracking-by-detection system. For detection a computationally demanding global CNN detector is applied to key-frames of each view interdependently. To get fast detections between key-frames a local CNN detector is combined with classical motion prediction. You *et al.* [102] present Deep Multi-Camera Tracking (DMCT), an end-to-end multi-view tracking and detection pipeline, which focuses on real-time applications.

A related class of approaches addresses the problem of multi-view crowd counting. Ge *et al.* [42] propose a generative probabilistic approach, using binary foreground masks as input. While the method is related to POM [38], the authors propose a sampling based strategy based on the Reversible Jump Markov Chain Monte Carlo (RJMCMC) [43] sampler instead of mean-field variation inference. This yields the advantage that a more complex posterior distribution can be approximated, in particular the global optimization can be obtained in continuous location space. More recent work focuses on deep learning techniques for crowd counting. Zhang *et al.* [103] introduce an end-to-end multi-view CNN architecture to predict the density map on the ground plane of the scene. For multi-view fusion the authors propose different late and early fusion strategies and report comparative results on different publicly available data sets. Castellano *et al.* [21] propose an alternative CNN architecture, focusing on the special demands of crowd counting from unmanned aerial vehicles.

2.2.2 Depth-based Approaches

In contrast to the RGB methods mentioned above, only a few existing approaches rely on multiple depth images for people detection. In this section we will first discuss methods focusing on pedestrian detection from the frontal or profile view. Those approaches follow the same paradigm and apply a classical single-view people detector to each RGB-D view independently [20, 69, 89, 91]. Subsequently, we will discuss methods which explicitly focus on top-view people detection and tracking in multiple depth images [64, 96]. While the former group of approaches in general leverages RGB and depth data (RGB-D), the latter only makes use of the depth information.

In early research Saputra *et al.* [89] propose indoor people tracking using two Kinect v1 sensors[1]. For people detection and tracking the authors apply human segmentation and the skeleton tracking, obtained by the Kinect for Windows SDK v1[2], to each view independently.

Following these ideas, Sun *et al.* [91] also employ a network of Kinect v1 sensors for people detection. Similar to [89], the human skeleton tracking from the Kinect for Windows SDK v1 is applied to each single view independently. The Kinect for Windows SDK v1 can only keep track of the skeletons of two individuals at the same time. To overcome this limitation a skeleton interleaving strategy is proposed to handle up to six individuals. Kalman filter based tracking is applied to foot points of individuals present in a single view. To handle occlusions, a trajectory matching scheme is proposed, in order to fuse trajectories from the same person (obtained by different views) to a single, global trajectory.

Munaro *et al.* [69] introduce the OpenPTtrack framework for people tracking in an RGB-D camera network. The project aims for an integrated, easy-to-use software system, including different camera calibration algorithms as well as sensor modalities. The two-stage people detection algorithm is performed on each view independently. First, a depth-based clustering is applied on the point cloud data. Second, for each cluster the

[1] The Microsoft Kinect v1 sensor is an active consumer RGB-D camera. The depth information is obtained by an actively emitted pattern in near-infrared, while an infrared camera captures the reflection of the pattern (structured light).

[2] https://www.microsoft.com/en-us/download/details.aspx?id=40278, accessed 31.03.2021

corresponding image region in the RGB image is fed into a HOG-based people detector. The resulting detections obtained by each sensor are gathered by a centralized tracking node and a typical tracking-by-detection scheme is applied.

Carraror *et al.* [20] propose an approach for human body pose estimation in a network of RGB-D sensors. To obtain a 3D skeleton, CNN-based pose estimation [99] is applied to the RGB images of each single-view. Incorporating the available depth data results in multiple 3D skeletons for each view. The single view 3D skeletons are further processed by a global node to solve the data association problem between the individual 3D skeletons, providing multi-view pose estimation.

In contrast to the aforementioned approaches, Tseng *et al.* [96] present an indoor people detection system based on multiple 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 detection a hemiellipsoidal head model is employed to take advantage of the discriminative height difference near the head contour of a human. In contrast to previously mentioned methods, the authors rely on depth data only. Furthermore, the detection is applied to the fused depth image, leveraging the full multi-view image evidence. In [64] the authors extend their approach, and focus on people tracking in a hybrid overlapping and non-overlapping camera network setup.

2.3 Integration of Present Thesis

In this section we will classify the contribution of the present work with respect to the aforementioned literature. As pointed out in Sect. 2.2.1 the vast majority of approaches rely on monocular video cameras but not on depth sensors for multi-view people detection. While in this thesis we address a different setup (top-view people detection with depth sensors), our work is methodically strongly related to multi-view approaches following the idea of generative modeling. Namely, these are POM, introduced by

Fleuret *et al.* [38], crowd detection with an MCMC sampler proposed by Ge *et al.* [42] and sparsity driven people localization introduced by Alahi *et al.* [6]. While, compared to the present work, the two latter approaches use very different methods for inference, our work is methodically most similar to [38]. Although we share a comparable mean-field inference strategy, there are several distinctions:

(i) our approach uses depth images as evidence and is therefore able to make use of a generative scene model based on a 3D shape model;

(ii) we introduce a novel strategy to approximate the final mean-field update expectation by making use of geometric scene knowledge, a pre-computed visual dictionary and a weighted asymmetric image similarity;

(iii) we propose to incorporate the temporal context jointly in the mean-field optimization to improve the detection performance.

According to the sensor modality our work belongs to the category of depth based multi-view approaches. In contrast to our work the vast majority of approaches in this category [20, 69, 89, 91] are based on RGB-D data and focus on the classical frontal or profile view. Besides, all aforementioned approaches apply single-view detection algorithms independently on each view and merge the local results to obtain global multi-view detections. Due to the single-view approach, they do not take advantage of the full multi-view information and heavily rely on the specific limitations of the employed people detector (e. g. viewpoint dependence).

To the best of our knowledge the only approach explicitly addressing the problem of indoor top-view people detection in multiple depth images is the work by Tseng *et al.* [96]. While our work addresses the same problem, their approach is methodically quite different. The authors fuse multiple depth views in a virtual top-view depth image and apply a discriminatively trained detector. While the proposed early fusion strategy successfully leverages the multi-view observations, evidenced by strong detection performance, their approach is focused on high quality depth data obtained by active depth sensors. In contrast, our method is particularly suitable for low resolution passive stereo sensors. Furthermore, our method approximates the joint probability distribution of people present in the scene, while [96] provide a MAP point estimate.

To summarize, methodically our work is highly inspired by Fleuret *et al.* [38], while from an application perspective it is most related to the work of Tseng *et al.* [96]. In contrast to recent data-driven CNN architectures [20, 23, 61, 67, 95] our method requires no training data and the detection confidence can be quantified more precisely by approximating the posterior distribution.

3 Background

In this chapter we will set the methodical foundations for the present thesis. In Sect. 3.1 we will outline the general idea of variational inference. In particular, we focus on the mean-field approximation which serves as the basis of the inference strategy proposed in this thesis. Based on this, we will discuss the application of mean-field variational inference to the problem of multi-view people detection in binary foreground images, referred to as probabilistic occupancy map (Sect. 3.2).

3.1 Mean-Field Variational Inference

A major challenge of probabilistic inference is that in many real world applications the posterior distribution over some symbolic latent variables $\mathcal{X}$ given the observations $\mathcal{O}$,

$$p(\mathcal{X} \mid \mathcal{O}) = \frac{p(\mathcal{X}, \mathcal{O})}{p(\mathcal{O})}\,, \tag{3.1}$$

is hard to compute or even intractable. Although it is possible for typical probabilistic models to compute the joint distribution $p(\mathcal{X}, \mathcal{O})$, the *evidence* or *marginal data likelihood*

$$p(\mathcal{O}) = \int p(\mathcal{X}, \mathcal{O})\, d\mathcal{X} \tag{3.2}$$

contains an integral (or a sum in case of discrete latent variables) over the full state space induced by $\mathcal{X}$, which is often intractable. The main idea of variational inference (VI) is to find a tractable proxy distribution

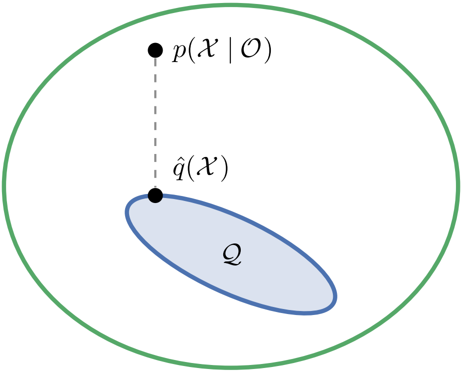


Figure 3.1 Schematic illustration of the variational inference objective. The green shape represents the space of all probability distributions, whereas $\mathcal{Q}$ represents a subset of possible proxy distributions (also referred to as q-family). The variational objective is to find a proxy distribution $\hat{q} \in \mathcal{Q}$ which is, with respect to some similarity measure, as close to the true posterior distribution $p(\mathcal{X} \mid \mathcal{O})$ as possible.

$q(\mathcal{X})$ which is, with respect to some similarity measure, as close to the true posterior $p(\mathcal{X} \mid \mathcal{O})$ as possible (see Fig. 3.1). The proxy distribution $q(\mathcal{X})$ can in general be an arbitrary probability distribution and is not restricted to some particular parametric form. In contrast to parameter estimation methods such as *maximum-likelihood estimation* (ML) or *maximum a posteriori estimation* (MAP), VI is a more general framework, since it allows to find the optimal distribution q by defining the objective as a functional rather than optimizing the parameters of a distribution explicitly.

The variational methods used in this work are part of the mean-field theory [74], which has its origin in statistical physics. The mean-field theory is applied in many scientific areas, in particular in statistical field theory, for example to understand ferromagnetism (cf. Ising model [25, 49]). In the last decades these methods haven been refined by the statistics and machine learning community [15, 28, 40, 50, 97] to make it applicable for general probabilistic inference. In this work we stick to the notation and conventions commonly used in the machine learning literature [11, 70].

To derive the objective for the optimal proxy distribution q one has to take further assumptions: First, one has to define a similarity measure between the true posterior distribution $p(\mathcal{X} \mid \mathcal{O})$ and the proxy distribution $q(\mathcal{X})$. The most popular choice is to use the Kullback-Leibler divergence (KL divergence, see Sect. 3.1.1). Following this assumption in Sect. 3.1.2, the general objective for KL variational inference is derived. Second, one has to take some assumptions on the structure of the proxy distribution q. In this work we use the so called *mean-field assumption*, which assumes that q factorizes over its marginal distributions. In virtue of this assumption it is possible to derive an optimization scheme, where the marginal probabilities $q(x'_i)$ are updated iteratively based on the current state of all other latent variables $\{x'_j : i \neq j\}$, see Sect. 3.1.3.

3.1.1 Kullback-Leibler Divergence

The Kullback-Leibler divergence [60], also known as relative entropy [26, p.19], has its origin in information theory. In the context of probability theory it is often interpreted as a measurement of dissimilarity between two probability distributions. Let p, q be the probability density functions of a continuous random variable, then the KL divergence is given as

$$\mathrm{KL}(p \parallel q) = \int p(\mathcal{X}) \log \left(\frac{p(\mathcal{X})}{q(\mathcal{X})} \right) d\mathcal{X} \,, \tag{3.3}$$

which can also be written as

$$\mathrm{KL}(p \parallel q) = -\int p(\mathcal{X}) \log \left(\frac{q(\mathcal{X})}{p(\mathcal{X})} \right) d\mathcal{X} \,. \tag{3.4}$$

For discrete probability distributions one has to replace the integral by the corresponding sum. The KL divergence is only defined if $q(\mathcal{X}) = 0$ implies $p(\mathcal{X}) = 0$. However, a common convention in the literature is that the KL divergence is set to infinity $\mathrm{KL}(p \parallel q) = \infty$ if there is any $\mathcal{X}$ such that $q(\mathcal{X}) = 0 \wedge p(\mathcal{X}) > 0$ (cf. [26, p. 19]). It can be proven that the KL divergence is non-negative $\mathrm{KL}(p \parallel q) \geq 0$ for any p, q and is zero if and only if $p = q$ (cf. [26, p. 28] and [11, p. 170]). It is important to notice that the KL divergence is asymmetric and in consequence does not fulfill the requirements of a distance metric.

An alternative perspective on the KL divergence follows trivially by applying the definition of expectation to (3.3), given as

$$\begin{aligned} \mathrm{KL}(p \parallel q) &= \left\langle \log \frac{p(\mathcal{X})}{q(\mathcal{X})} \right\rangle_{p(\mathcal{X})} && (3.5) \\ &= \langle \log p(\mathcal{X}) - \log q(\mathcal{X}) \rangle_{p(\mathcal{X})} , && (3.6) \end{aligned}$$

where $\langle \cdot \rangle_{p(\mathcal{X})}$ denotes the expectation with respect to the distribution $p(\mathcal{X})$.

3.1.2 KL Variational Inference

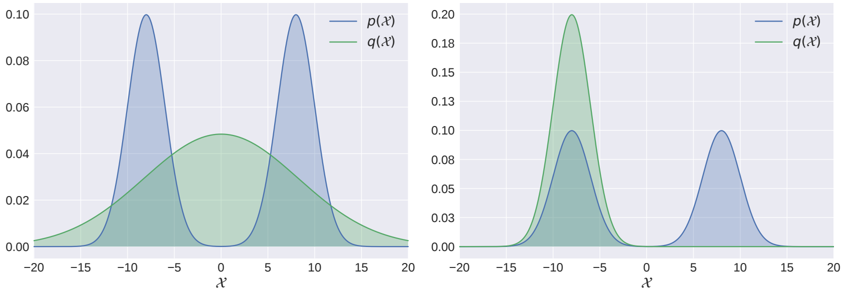
Since the KL divergence is asymmetric, the order of the arguments does have an impact on the variational objective. In the literature $\mathrm{KL}(p \parallel q)$ is referred to as forward KL divergence and $\mathrm{KL}(q \parallel p)$ as reverse KL divergence respectively. Inspecting the definition of the forward KL divergence

$$\mathrm{KL}(p(\mathcal{X} \mid \mathcal{O}) \parallel q(\mathcal{X})) = \int p(\mathcal{X} \mid \mathcal{O}) \log \left(\frac{p(\mathcal{X} \mid \mathcal{O})}{q(\mathcal{X})} \right) d\mathcal{X} , \tag{3.7}$$

one can see that the forward KL divergence is infinity if $q(\mathcal{X}) = 0$ and $p(\mathcal{X} \mid \mathcal{O}) > 0$. Thus, if using the forward KL divergence as an objective, $q(\mathcal{X})$ gets forced to be non-zero if $p(\mathcal{X} \mid \mathcal{O}) > 0$. On the other hand, the reverse KL divergence

$$\mathrm{KL}(q(\mathcal{X}) \parallel p(\mathcal{X} \mid \mathcal{O})) = \int q(\mathcal{X}) \log \left(\frac{q(\mathcal{X})}{p(\mathcal{X} \mid \mathcal{O})} \right) d\mathcal{X} , \tag{3.8}$$

has the opposite effect. It approaches infinity if $p(\mathcal{X} \mid \mathcal{O}) = 0$ and $q(\mathcal{X}) > 0$. Following the same argument as above, $q(\mathcal{X})$ gets forced to be zero for $p(\mathcal{X} \mid \mathcal{O}) = 0$. One can conclude that when fitting a distribution q



(a) $\mathrm{KL}(p \parallel q)$

(b) $\mathrm{KL}(q \parallel p)$

Figure 3.2 Comparison of the results when fitting a single univariate normal distribution q to a univariate bimodal distribution p with respect to the forward (a) and the reverse (b) KL divergence. The exemplary "true" distribution $p(\mathcal{X})$ (blue curve) is given as a mixture of Gaussian $p(\mathcal{X}) = 0.5(\mathcal{N}(\mathcal{X} \mid -\mu, \sigma)) + \mathcal{N}(\mathcal{X} \mid \mu, \sigma))$ with $\mu = 8, \sigma = 2$. (a) shows the optimal fitting normal distribution q with respect to the forward KL divergence, given by the moments of $p(\mathcal{X})$. (b) shows q with respect to the reverse KL divergence respectively. In contrast to (a), the optimal choice of q is not given in closed form for the reverse KL divergence. In this example the optimum is to locally fit q to one of the modes of p (example based on [11, p. 619 ff.]).

to a distribution p with respect to the forward KL divergence, q is in general overestimating the support of p. In contrast, using the reverse KL divergence is potentially underestimating the support of the true posterior p.

Studying Fig. 3.2 and following the arguments in [70, p. 733 f.] one can see that in a scenario where p is a multimodal and q a unimodal distribution, using the forward KL divergence is problematic. Since q is forced to be a unimodal distribution, the mode of q is somewhere between the modes of p, which is a region where p has rather low probability density.

Another aspect is the tractability of the resulting optimization problem. Using the forward KL divergence one has to compute the expectation with respect to the intractable distribution $p(\mathcal{X})$ which is in general hard to compute. However, using the reverse KL divergence yields the advantage that expectation is computed with respect to the simpler proxy distribution

$q(\mathcal{X})$. Depending on the choice of the q-family (thus the structure of q) this leads to a feasible optimization problem (see Sect. 3.1.3). Therefore, it is common to use the reverse KL divergence to formulate the variational objective

$$\hat{q}(\mathcal{X}) = \operatorname*{arg\,min}_{q \in \mathcal{Q}} \mathrm{KL}(q(\mathcal{X}) \parallel p(\mathcal{X} \mid \mathcal{O})) \,, \tag{3.9}$$

with the KL divergence

$$\begin{aligned} \mathrm{KL}(q(\mathcal{X}) \parallel p(\mathcal{X} \mid \mathcal{O})) &= \langle \log q(\mathcal{X}) - \log p(\mathcal{X} \mid \mathcal{O}) \rangle_{q(\mathcal{X})} & (3.10) \\ &= \langle \log q(\mathcal{X}) \rangle_{q(\mathcal{X})} - \langle \log p(\mathcal{X} \mid \mathcal{O}) \rangle_{q(\mathcal{X})} \,. & (3.11) \end{aligned}$$

Expanding the expectation over the conditional distribution we can unveil the dependence of the evidence $p(\mathcal{O})$

$$\begin{aligned} \langle \log p(\mathcal{X} \mid \mathcal{O}) \rangle_{q(\mathcal{X})} &= \langle \log p(\mathcal{X}, \mathcal{O}) \rangle_{q(\mathcal{X})} - \langle \log p(\mathcal{O}) \rangle_{q(\mathcal{X})} & (3.12) \\ &= \langle \log p(\mathcal{X}, \mathcal{O}) \rangle_{q(\mathcal{X})} - \log p(\mathcal{O}) \,. & (3.13) \end{aligned}$$

Inserting into (3.11) we can rewrite the KL divergence as

$$\begin{aligned} \mathrm{KL}(q(\mathcal{X}) \parallel p(\mathcal{X} \mid \mathcal{O})) = &\underbrace{\langle \log q(\mathcal{X}) \rangle_{q(\mathcal{X})} - \langle \log p(\mathcal{X}, \mathcal{O}) \rangle_{q(\mathcal{X})}}_{\mathrm{KL}(q(\mathcal{X}) \parallel p(\mathcal{X}, \mathcal{O}))} \\ &+ \log p(\mathcal{O}) \,. \end{aligned} \tag{3.14}$$

Since $\log p(\mathcal{O})$ does not depend on $\mathcal{X}$ it can be treated as a constant and does not affect the solution of the objective. Thus minimizing the original KL divergence (3.11) is equal to minimizing the KL divergence $\mathrm{KL}(q(\mathcal{X}) \parallel p(\mathcal{X}, \mathcal{O}))$. To get further insights we rearrange (3.14) to

$$\begin{aligned} \log p(\mathcal{O}) = \mathrm{KL}(q(\mathcal{X}) \parallel p(\mathcal{X} \mid \mathcal{O})) - \langle \log q(\mathcal{X}) \rangle_{q(\mathcal{X})} \\ + \langle \log p(\mathcal{X}, \mathcal{O}) \rangle_{q(\mathcal{X})} \,. \end{aligned} \tag{3.15}$$

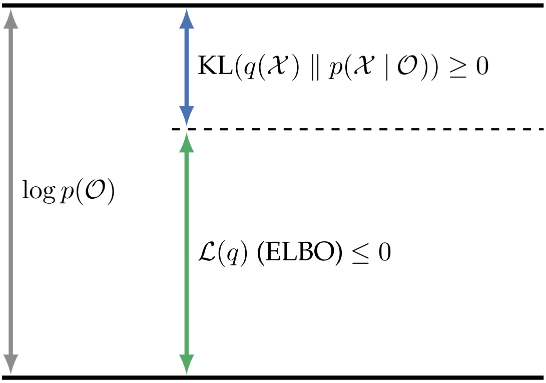

$\mathrm{KL}(q(\mathcal{X}) \parallel p(\mathcal{X} \mid \mathcal{O})) \geq 0$
$\log p(\mathcal{O})$
$\mathcal{L}(q)$ (ELBO) ≤ 0

Figure 3.3 Graphical illustration of (3.15). Maximizing the ELBO is equivalent to minimizing the KL divergence $\mathrm{KL}(q(\mathcal{X}) \parallel p(\mathcal{X} \mid \mathcal{O}))$. The KL divergence can also be interpreted as the error between the lower bound $\mathcal{L}(q)$ and the actual evidence $\log p(\mathcal{O})$. Figure inspired by Bishop [14, Fig. 9.11].

Since $\mathrm{KL}(\cdot \parallel \cdot) \geq 0$, (3.15) implies a lower bound $\mathcal{L}(q)$ for the marginal data likelihood

$$\log p(\mathcal{O}) \geq \underbrace{-\langle \log q(\mathcal{X}) \rangle_{q(\mathcal{X})} + \langle \log p(\mathcal{X}, \mathcal{O}) \rangle_{q(\mathcal{X})}}_{\mathcal{L}(q)} . \tag{3.16}$$

In Fig. 3.3 a graphical illustration of (3.15) and (3.16) is given. In the machine learning literature the lower bound $\mathcal{L}(q)$ is referred to as *evidence lower bound (ELBO)*. Notice that the definition of the ELBO is related to the concept *free energy* in statistical thermodynamics, therefore $-\mathcal{L}(q)$ is also referred to as *variational free energy* (cf. [70, Sect. 21.2.1]). As a consequence of definition (3.4) the ELBO can also be written as

$$\mathcal{L}(q) = -\mathrm{KL}(q(\mathcal{X}) \parallel p(\mathcal{X}, \mathcal{O})) . \tag{3.17}$$

Maximizing the ELBO $\mathcal{L}(q)$ with respect to q has two relevant aspects:

(i) In face of (3.15) and (3.17) one can see that the original objective (3.9) is equivalent to the objective

$$\hat{q} = \arg\max_{q} \mathcal{L}(q) \,. \tag{3.18}$$

Thus maximizing the ELBO (3.16) leads to the desired approximation of the conditional distribution $p(\mathcal{X} \mid \mathcal{O})$.

(ii) As a direct consequence of the definition (3.16), maximizing $\mathcal{L}(q)$ yields a lower bound on the marginal data likelihood $p(\mathcal{O})$. This is particularly useful for Bayesian model selection (cf. [11, p. 619]).

3.1.3 Mean-Field Approximation

Having defined the variational objective in the previous section as maximizing the ELBO $\mathcal{L}(q)$ with respect to some distribution $q(\mathcal{X})$, one still has to restrict the set $\mathcal{Q}$ of possible distributions $q \in \mathcal{Q}$ to make the optimization feasible. In general, it is possible to do so by (i) assuming that q is of a specific functional or parametric form; (ii) assuming that q has a specific structure, i.e. defining how q factorizes over the latent variables $\mathcal{X}$. While the first assumption (i) is more restrictive, (ii) is more general and exploits the full power of the variational approach. Notice that it is also possible to combine both assumptions.

In this work we apply the so called *mean-field assumption*. While in general this assumption allows factorizing $q(\mathcal{X})$ over arbitrary disjoint partitions of latent variables [14, p. 464], in this work we apply the mean-field assumption in its simplest form[1]. We assume that $q(\mathcal{X})$ is a fully factorized distribution, given as product over its marginal probabilities

$$q(\mathcal{X}) = \prod_{i=1}^{r} q_i(x_i') \,, \tag{3.19}$$

[1] Sometimes referred to as *naive mean-field assumption* in the literature [11, p. 623].

where r is the number of latent variables. Even though this is a strong simplification on the structure of possible distributions q, there are no restrictions on the functional form of the individual distributions q_i. In practice, it turns out that this simple factorial structure leads to a tractable variational objective for many probabilistic models.

In order to deduce the mean-field update equations one needs to define a probability distribution over $\mathcal{X}$ except of one single element $x_i' \in \mathcal{X}$. Therefore, let

$$q\big(\mathcal{X} \setminus x_i'\big) = \prod_{j=1:j\neq i}^{r} q_j(x_j') \tag{3.20}$$

denote the mean-field distribution excluding the element x_i'.

3.1.3.1 Mean-Field Update Equations

Using the results from the previous section the general objective is to maximize the ELBO with respect to the function q,

$$\hat{q} = \arg\max_{q} \mathcal{L}(q)\,. \tag{3.21}$$

Applying the mean-field assumption, one possible optimization technique is to apply a coordinate descent strategy, where one marginal distribution q_i gets updated, while the other variables stay fixed. To derive the corresponding update equations one can isolate the dependency of one single q_i. Following the detailed derivation in appendix A.1, it follows that the ELBO $\mathcal{L}(q_i)$ for a single q_i can be written as a negative KL divergence

$$\mathcal{L}(q_i) = -\mathrm{KL}\big(q_i(x_i') \parallel \tilde{p}_i(\mathcal{X}, \mathcal{O})\big) + \mathrm{const}\,, \tag{3.22}$$

with the distribution

$$\tilde{p}_i(\mathcal{X}, \mathcal{O}) = \frac{1}{Z_i} \exp\Big(\langle \log p(\mathcal{X}, \mathcal{O})\rangle_{q(\mathcal{X}\setminus x_i')}\Big)\,, \tag{3.23}$$

where Z_i is the partition function and the notation $\langle \cdot \rangle_{q(\mathcal{X} \setminus x'_i)}$ refers to the expectation with respect to the distribution $q(\cdot)$ for all $\{x'_j : i \neq j\}$:

$$\langle \log p(\mathcal{X}, \mathcal{O}) \rangle_{q(\mathcal{X} \setminus x'_i)} = \int \log p(\mathcal{X}, \mathcal{O}) \prod_{j=1:j \neq i}^{r} q_j(x'_j)\, d\mathcal{X} \backslash x'_i \,. \qquad (3.24)$$

Up to a constant term, $\mathcal{L}(q_i)$ (3.22) is equivalent to a negative KL divergence between the desired distribution $q_i(x'_i)$ and a distribution proportional to $\exp\Big(\langle \log p(\mathcal{X}, \mathcal{O}) \rangle_{q(\mathcal{X} \setminus x'_i)}\Big)$. According to (3.22) we can maximize the ELBO for one single distribution $q_i(x'_i)$ by minimizing $\mathrm{KL}(q_i \parallel \tilde{p}_i)$. It trivially follows that the optimal $\hat{q}_i(x'_i)$ with respect to the objective

$$\hat{q}_i = \arg\max_{q_i} \mathcal{L}(q_i)\,, \qquad (3.25)$$

satisfies $\hat{q}_i = \tilde{p}_i$. This finally leads to the so called *general mean-field equations* [11, p. 625 ff.]

$$\begin{aligned} q_i(x'_i) &= \frac{1}{Z_i} \exp\Big(\langle \log p(\mathcal{X}, \mathcal{O}) \rangle_{q(\mathcal{X} \setminus x'_i)}\Big) \\ &\propto \exp\Big(\langle \log p(\mathcal{X}, \mathcal{O}) \rangle_{q(\mathcal{X} \setminus x'_i)}\Big)\,, \end{aligned} \qquad (3.26)$$

for all marginal distributions q_i with $i \in \{1, \ldots, r\}$. Fixing all other latent variables $\mathcal{X} \setminus x'_i$, this equation updates a single distribution $q_i(x_i)$ with respect to the current mean-field state $q\big(\mathcal{X} \setminus x'_i\big)$. This is a general result of the mean-field assumption and does not hinge on a specific parametric distribution family of q_i. In practice the resulting parametric form of the marginal distributions q_i and the tractability of the mean-field update equations depend in particular on the probabilistic model $p(\mathcal{X}, \mathcal{O})$. Thus one essential part of mean-field variational inference is to define a factorization for q and a probabilistic model $p(\cdot)$ leading to tractable mean-field update equations. Although now we have an individual equation for each marginal distribution q_i, in general we cannot write this as a closed form solution since each equation depends on the states of all other marginal distributions $q\big(\mathcal{X} \setminus x_i'\big)$. This commonly leads to an iterative update scheme, where one distribution q_i is updated with respect to the current state $q\big(\mathcal{X} \setminus x_i'\big)$, which we will discuss in the following section. A detailed discussion on the general mean-field update equations can be found in [70, p. 736 ff.], [14, p. 465 ff.] and [15].

3.1.4 Coordinate Ascent Mean-Field Variational Inference

In the previous sections we recast the general probabilistic inference problem to an optimization problem, which is finally given by the general mean-field equations (3.26) for each marginal distribution q_i. A popular algorithm to solve this optimization problem is referred to as *Coordinate Ascent Variational Inference* (CAVI) [15]. In order to decouple the dependence of the marginal distributions, CAVI iteratively updates one marginal distribution q_i after the other, while all other distributions $q_j : j \neq i$ stay fixed. Hence, each q_i is updated with respect to the current mean-field state $q\big(\mathcal{X} \setminus x_i'\big)$ as presented in Algorithm 1. In practice it is not necessarily

Algorithm 1 Coordinate Ascent Mean-Field Variational Inference (CAVI)

```
procedure OPTIMIZE Q (Data: $\mathcal{O}$, Model: $p(\mathcal{X}, \mathcal{O})$)
    $q(\mathcal{X}) \leftarrow$ init()                                   ▷ init mean-field
    while $\mathcal{L}(q)$ has not converged do
        for all $i \in \{1, \dots, r\}$ do                          ▷ iterate over marginals
            $q_i \leftarrow \exp \langle \log p(\mathcal{X}, \mathcal{O}) \rangle_{q(\mathcal{X} \setminus x_i')}$

            $q(\mathcal{X}) \leftarrow q_i$                       ▷ asynchronous update of MF state
        end for
        $\mathcal{L}(q) \leftarrow \langle \log p(\mathcal{X}, \mathcal{O}) \rangle_{q(\mathcal{X})} - \langle \log q(\mathcal{X}) \rangle_{q(\mathcal{X})}$    ▷ compute ELBO
    end while
    return $\hat{q}$
end procedure
```

needed to use the ELBO $\mathcal{L}(q)$ as a termination condition of the optimization. It is often sufficient to monitor the convergence behavior of the marginal distributions q_i.

The CAVI algorithm relies on asynchronous mean-field updates (3.26), which means that the individual $q_i(x_i)$'s are updated sequentially in each iteration. It can be proven that updating each q_i asynchronously gradually improves the approximation for each q_i (cf. [11, p. 626]). In contrast, it is also possible to execute the mean-field updates synchronously. In a synchronous update iteration all the q_i are updated simultaneously, using the same mean-field state obtained from the previous iteration. While synchronous mean-field updates can be easily parallelized, the convergence guarantees from the asynchronous updates do not hold anymore. In general synchronous updates can lead to oscillating effects during the optimization. In Sect. 6.4.1.2 we report an example of this effect, in particular illustrated in Fig. 6.7.

3.2 Probabilistic Occupancy Map (POM)

In 2008 Fleuret *et al.* [38] introduce the probabilistic occupancy map (POM). For almost a decade, POM served as the state-of-the-art method for multi-view people detection in RGB images. Although recent state-of-art methods such as Deep Occlusion [10] outperform classical POM, Deep Occlusion still partially relies on the generative model introduced by POM. Since the present work is inspired by POM, in this section we will discuss the technical details of POM and reveal the relations to this thesis.

POM solves the problem of multi-view people detection as an inverse problem. Multiple overlapping foreground-segmented binary images are used as input and compared against synthesized binary images obtained by a generative model. Each individual person is expressed by an axially parallel rectangular box (see Fig. 3.4(a)) in order to generate synthetic foreground images. For inference the KL divergence between the true posterior distribution and a fully factorized distribution is minimized by an iterative optimization algorithm. Even though in the original work [39, 38] the authors do not refer to the method of mean-field variational inference, the update equations derived in [38, Eq. 25] satisfy the general mean-field equations (3.26). To reveal the connection of POM to the present work we use the notation and terminology of mean-field variational inference introduced in Sect. 3.1 as well as the notation of the discrete probabilistic scene model introduced in Sect. 4.4.

3.2.1 Probabilistic Model

Let $\mathbf{b} = (b_1, \ldots, b_C)^\intercal$ be the input observations, given as foreground segmented binary images obtained from C cameras at one time step. Since POM does not depend on the temporal context we will omit the time index in this section to keep the notation clear. To define the occupancy grid, the plane at ground level is discretized into a regular 2D-grid of n locations. Each grid location u_i will be assigned a realization x_i of a Bernoulli random variable $X_i \sim \mathcal{B}(\mu_i)$, where μ_i denotes the probability of a person present at location u_i. The occupancy map is represented as the vector $\mathbf{x} = (x_1, \ldots, x_n)^\intercal \in \{0, 1\}^n$ and is also referred to as a (discrete) scene configuration in this thesis (cf. Sect. 4.4). The joint distribution over the observations and scene configuration is given as

$$p(\mathbf{b}, \mathbf{x}) = p(\mathbf{b} \mid \mathbf{x}) p(\mathbf{x})\,. \tag{3.27}$$

Fleuret *et al.* [38] employ two independence assumptions to make the distribution (3.27) tractable (cf. [38, Sect. 5.1]).

First it is assumed that the prior $p(\mathbf{x})$ factorizes as

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i)\,. \tag{3.28}$$

This implies the assumption of individuals moving around independently of each other. While in reality this assumption does not hold e. g. since individuals are in general keeping a certain inter-person distance to each other (cf. Sect. 4.3.2), it is a common assumption to keep the posterior distribution tractable.

Second, for POM it is assumed that the statistical dependence between views is only given by the people present in the scene. Thus, if the scene configuration $\mathbf{x}$ is fixed, the views are assumed to be conditionally independent:

$$p(\mathbf{b} \mid \mathbf{x}) = \prod_{c=1}^{C} p(b_c \mid \mathbf{x}) \,. \tag{3.29}$$

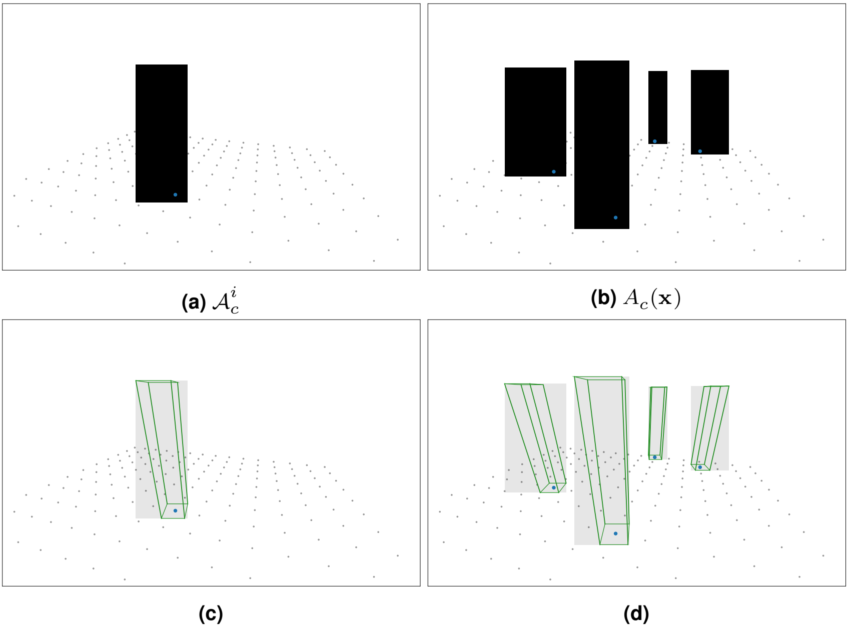
Although this assumption does not cover global effects, such as illumination changes etc., it can be justified by considering that b_c are foreground segmented images, while people are assumed to be the only moving objects in the scene.

3.2.1.1 Generative Scene Model

To be able to define the likelihood term $p(b_c \mid \mathbf{x})$ Fleuret *et al.* [38] propose a generative model $A_c(\mathbf{x})$, which maps a scene configuration $\mathbf{x}$ to a synthetic binary image in the perspective of camera c. It is assumed that the cameras used are extrinsically and intrinsically calibrated. To generate a synthetic image $A_c(\mathbf{x})$, fixed sized cuboids are placed in the 3D scene depending on the provided scene configuration $\mathbf{x}$. Each cuboid is rendered into all views using the given extrinsic and intrinsic camera parameters. For each rendered cuboid the axis-aligned bounding box is computed and the corresponding pixels are classified as foreground pixels in the resulting synthetic images. In Fig. 3.4 the synthetic image generation process is illustrated in detail.

Let $\mathcal{A}_c^i$ be a binary synthetic image with only a rectangle placed at the grid location with index i, where all pixels inside the rectangle are equal to 1 (Fig. 3.4(a)). The synthetic image generation is formally given as

$$A_c(\mathbf{x}) = \cup_{i=1}^{n} x_i \mathcal{A}_c^i \,, \tag{3.30}$$



(a) $\mathcal{A}_c^i$

(b) $A_c(\mathbf{x})$

(c)

(d)

Figure 3.4 Illustration of generative scene model producing binary images in the perspective of camera c. Each gray dot represents the center of a grid cell on the ground plane assigned with a Bernoulli random variable X_i (for visualization only). Each black rectangle in (a) and (b) represents an individual present at a particular grid cell. (a) shows a synthetic binary image with exactly one individual placed at grid location i. (b) shows a generated synthetic image corresponding to an exemplary scene configuration $\mathbf{x}$ with four non-zero entries. The dimensions of the rectangles are given as the axis-aligned minimum bounding box of a cuboid, roughly approximating the shape of a person, as shown in (c) and (d). Figures inspired by [38, Fig. 4,6]

where the union $\cup$ is defined as the binary image union (pixel-wise logical OR). An exemplary synthetic image $A_c(\mathbf{x})$ is given in Fig. 3.4(b). Based on this generative model, the likelihood for a single observation b_c is given as

$$p(b_c \mid \mathbf{x}) \propto \exp\big(-\delta_{\text{pom}}(b_c, A_c(\mathbf{x}))\big) , \tag{3.31}$$

where $\delta_{\text{pom}}(b_c, A_c(\mathbf{x})))$ is a pseudo distance function between an observed image b_c and a synthetic image $A_c(\mathbf{x})$. Let $a, b \in [0,1]^{W \times H}$ be images with the dimension $W \times H$ pixels, the pseudo distance function is given as

$$\delta_{\text{pom}}(b, a) = \xi^{-1} \cdot \frac{\|b \odot (1-a) + (1-b) \odot a\|_1}{\|a\|_1}\,, \tag{3.32}$$

with $\|\cdot\|_1 = \|\text{vec}(\cdot)\|_1$ defined as the sum of all pixels (element-wise L1 vector norm). The parameter ξ can be interpreted as a pseudo standard deviation, controlling the reliability of the measurements. Finally, the desired posterior distribution is given as

$$\begin{align} p(\mathbf{x} \mid \mathbf{b}) &= \frac{1}{Z} \prod_{c=1}^{C} p(b_c \mid \mathbf{x}) \prod_{i=1}^{n} p(x_i) \tag{3.33} \\ &\propto \prod_{c=1}^{C} \exp\bigl(-\delta_{\text{pom}}(b_c, A_c(\mathbf{x}))\bigr) \prod_{i=1}^{n} p(x_i)\,. \tag{3.34} \end{align}$$

3.2.2 Inference

Due to the high dimensional latent space $\{0,1\}^n$ the posterior distribution (3.33) is intractable. Fleuret *et al.* [38] propose to apply mean-field variational inference (cf. Sect. 3.1) to approximate the posterior distribution $p(\mathbf{x} \mid \mathbf{b})$ with a simpler proxy distribution $q(\mathbf{x}) = \prod_{i=1}^{n} q_i(x_i)$ (mean-field assumption). The objective is given as

$$\hat{q}(\mathbf{x}) = \arg\min_{q} \text{KL}(q(\mathbf{x}) \parallel p(\mathbf{x} \mid \mathbf{b}))\,. \tag{3.35}$$

According to the general mean-field equations (3.26) the optimal update for the marginal distributions q_i satisfy

$$q_i(x_i = 1) = \frac{1}{Z_i} \exp\Bigl(\langle \log p(\mathbf{b}, \mathbf{x} \mid x_i = 1) \rangle_{q(\mathbf{x} \setminus x_i)}\Bigr)\,, \tag{3.36}$$

with the partition function

$$Z_i = \sum_{s \in \{0,1\}} \exp\Big(\langle \log p(\mathbf{b}, \mathbf{x} \mid x_i = s) \rangle_{q(\mathbf{x} \setminus x_i)} \Big) \, . \tag{3.37}$$

Inserting the probabilistic model defined in (3.33) the update simplifies to

$$q_i(x_i = 1) = \left[1 + \exp\left(\tau_i + \sum_{c=1}^{C} \Psi_{c,i}^{\text{pom}} \right) \right]^{-1} , \tag{3.38}$$

with the prior term $\tau_i = \log \frac{1-p(x_i=1)}{p(x_i=1)}$ and the data expectation term for one camera given as

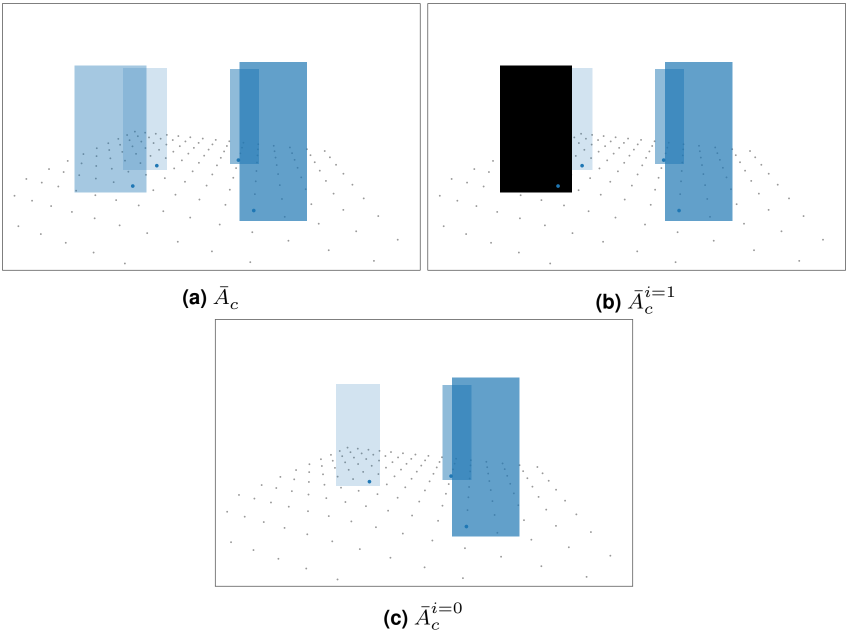
$$\begin{aligned} \Psi_{c,i}^{\text{pom}} &= \big\langle \delta_{\text{pom}}(b_c, A_c(\mathbf{x}|x_i{=}1)) \big\rangle_{q(\mathbf{x} \setminus x_i)} \\ &\quad - \big\langle \delta_{\text{pom}}(b_c, A_c(\mathbf{x}|x_i{=}0)) \big\rangle_{q(\mathbf{x} \setminus x_i)} \, . \end{aligned} \tag{3.39}$$

A detailed derivation of the POM update equations (3.38) is given in Appendix A.2. For more insights we additionally refer to Sect. 5.2.1, where the mean-field update equations are derived and interpreted for a similar posterior distribution. Note that although the derivation given in this section is different from the one given in the original POM paper [38], the resulting update equations for the distributions q_i are identical (cf. [38, Eq. 25]).

3.2.2.1 Approximation

Since the expectations in (3.39) are not tractable, the POM authors propose to approximate the expectations by only taking into account the expectation of the image generation process rather than the full image distance. For $s \in \{0, 1\}$ the approximation is given as (cf. [38, Eq. 26])

$$\big\langle \delta_{\text{pom}}(b_c, A_c(\mathbf{x}|x_i{=}s)) \big\rangle_{q(\mathbf{x} \setminus x_i)} \simeq \delta_{\text{pom}}(b_c, \langle A_c(\mathbf{x}|x_i{=}s) \rangle_{q(\mathbf{x} \setminus x_i)}) \, . \tag{3.40}$$



(a) $\bar{A}_c$

(b) $\bar{A}_c^{i=1}$

(c) $\bar{A}_c^{i=0}$

Figure 3.5 Schematic illustration of synthetic average images. (a) shows an exemplary average image for a distribution $q(\mathbf{x})$ with four non-zero marginal distributions $q_j(x_j)$. (b) conditional synthetic average image $\bar{A}_c^{i=1} = \langle A_c(\mathbf{x}|x_i{=}1)\rangle_{q(\mathbf{x}\setminus x_i)}$ where the black rectangle corresponds to the grid location i. (c) shows the corresponding conditional average image given that $x_i = 0$. The images (a) - (c) only differ at the pixel values in the rectangle related to the grid cell with index i. Figures based on [38, Fig. 6].

Due to the binary images the expectation over the synthetic images

$$\bar{A}_c = \langle A_c(\mathbf{x})\rangle_{q(\mathbf{x}\setminus x_i)} \tag{3.41}$$

is easy to compute and can be interpreted as an average image (see Fig. 3.5(a)). The unconditioned average image $\bar{A}_c$ is defined for every pixel (u, v) as (cf. [38, Eq. 30])

$$\bar{A}_c[u,v] = \langle A_c(\mathbf{x})[u,v] \rangle_{q(\mathbf{x})} = 1 - \prod_{j:\mathcal{A}_c^j[u,v]=1} (1 - q_j(x_j = 1)) \,, \quad (3.42)$$

where the index set $\{j \mid \mathcal{A}_c^j[u,v] = 1\}$ includes any grid index j for which the corresponding synthetic image $\mathcal{A}_c^j$ contains a foreground pixel at the image coordinate (u, v). The inverse probability in (3.42) ensures that pixels not influenced by any q_j are assigned to a probability of zero. The conditional synthetic average image for camera c with respect to the current mean-field state $q(\mathbf{x})$ and x_i forced to state $s \in \{0, 1\}$ is given as

$$\bar{A}_c^{i=s} = \langle A_c(\mathbf{x}|x_i = s) \rangle_{q(\mathbf{x} \setminus x_i)} \,. \quad (3.43)$$

Considering the state $s \in \{0, 1\}$, the average image $\bar{A}_c^{i=s}$ is obtained by forcing all pixels in $\bar{A}_c[u, v]$ effected by the state of x_i (all pixels of $\mathcal{A}_c^i$ which are equal to 1) to 0 or 1 respectively (see Fig. 3.5(b,c)).

This leads to the following two implications: (i) the conditional synthetic average image $\bar{A}_c^{i=s}$ is a function of $q(\mathbf{x} \setminus x_i)$; (ii) the conditioned average images $\bar{A}_c^{i=0}$ and $\bar{A}_c^{i=1}$ only differ from the average image $\bar{A}_c$ in the pixels effected by the rectangle $\mathcal{A}_c^i$ related to x_i. The final POM update equations are given as

$$q_i(x_i = 1) = \frac{1}{1 + \exp\Big(\tau_i + \sum_{c=1}^{C} \Big(\delta_{\text{pom}}(b_c, \bar{A}_c^{i=1}) - \delta_{\text{pom}}(b_c, \bar{A}_c^{i=0})\Big)\Big)} \,. \quad (3.44)$$

In order to estimate the distribution q, Fleuret *et al.* [38] propose an iterative update scheme. In contrast to CAVI (Sect. 3.1.4), which uses asynchronous mean-field updates, a synchronous update is used. This yields the advantage that the average image $\bar{A}_c$ needs only to be calculated once per iteration. Moreover, by exploiting integral images, the distance $\delta_{\text{pom}}(b_c, \bar{A}_c^{i=s})$ can be calculated in constant time. In total this leads to a real-time capable estimation of the marginal distributions $q_i(x_i)$. For a detailed explanation of the fast iterative update scheme used by POM we refer to the original paper [38, Sect. 5.4].

4 Probabilistic Model

In this chapter we introduce a probabilistic model for people detection in multiple depth images. First we declare basic prerequisites, covering aspects such as sensor network calibration and multi-view image evidence (Sect. 4.1). Hereinafter, we propose an abstract general probabilistic model (Sect. 4.2), which includes the generative scene model and the related modeling of the data likelihood. Based on the general model, we derive two different manifestations, sharing the same generative scene model and functional structure (e. g. conditional independence of the views). Namely, we put the general model in concrete terms for the continuous latent space (Sect. 4.3) and for the discrete latent space (Sect. 4.4), respectively. All the common aspects of the discrete and continuous model are covered by the abstract general model.

4.1 Prerequisites

4.1.1 Sensor Network Setup

While the methods proposed in this work can in general be applied to any kind of depth sensor network, we will take some initial assumptions on the setup of the depth sensor network in order to simplify further modeling.

First, it is assumed that the observed scene has one common plane at ground level, which we refer to as ground plane in this thesis. This assumption can be justified considering typical man-made environments, e. g. train stations, shopping malls or other indoor environments. Of course this assumption does not hold in some particular relevant scenarios, e. g.

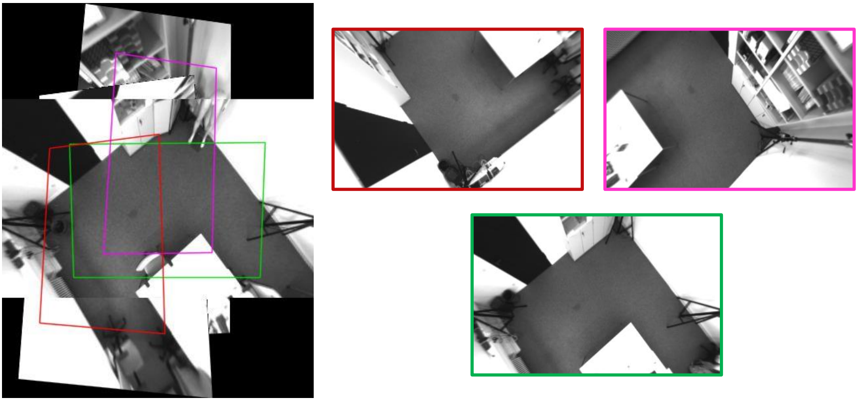


Figure 4.1 Stitched multi-sensor image based on marker-free calibration. The colored quadrilaterals in the left image represent the view frustums intersected with the plane one meter parallel to the ground level. The three images on the right show the individual sensor views with corresponding color encoding.

a staircase connected to a hallway. It is viable to combine the proposed method with a more general ground plane model, e. g. a composite of multiple planes, in order to handle such scenarios.

Second, we assume that the depth sensors are intrinsically and extrinsically calibrated in advance. If the sensors share overlapping fields of view in a way that there is at least one path connecting all sensors, the extrinsic calibration can be obtained by visual cues. Due to the available depth data, marker-free extrinsic calibration can be achieved in three steps: (i) for each sensor $\mathcal{S}_1, \ldots, \mathcal{S}_C$ the ground plane is estimated by a plane fit; (ii) one arbitrary sensor coordinate system is defined as the common world coordinate system; (iii) for the other sensors $\mathcal{S}_c$ the rigid body transformation to the common world coordinate system is obtained by the corresponding natural image features in the overlapping fields of view. Due to the available depth data, each natural image feature match implies a world point correspondence. For the rest of this work we define $\mathbf{P}_c$ as the projection matrix for each sensor $\mathcal{S}_c$, which maps a point from the common world coordinate system to the corresponding image coordinates of each sensor.

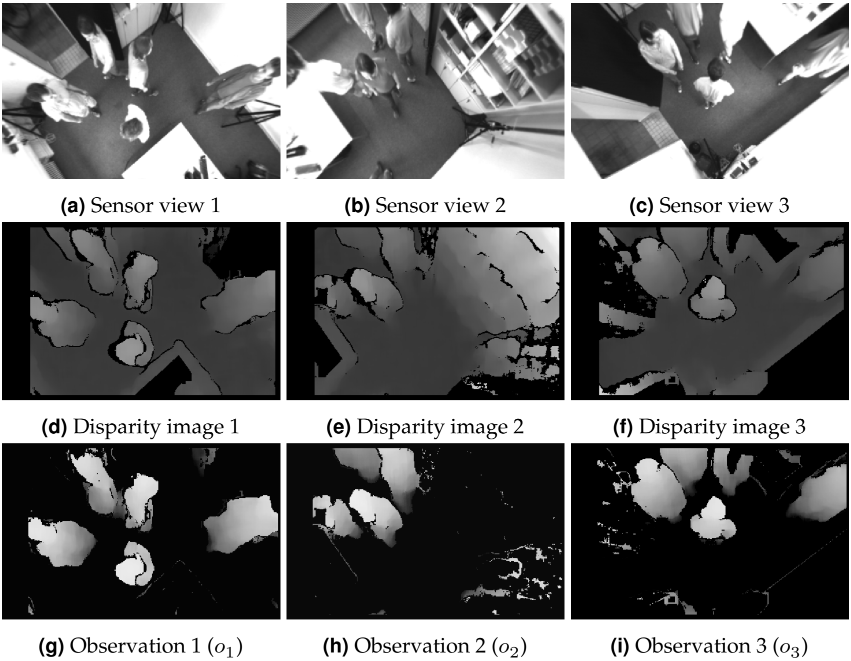
As a direct consequence of the two former assumptions, we describe the presence of people in the scene in ground plane world coordinates $(x, y) \in \mathbb{R}^2$. A particular constellation of individuals on the ground plane is referred to as a scene configuration. In this thesis scene configurations are given in two manifestations: a list of continuous plane coordinates (cf. Sect. 4.3) or as a discrete ground plane grid (cf. Sect. 4.4).

4.1.2 Observations

As image evidence, we use low-resolution depth images, obtained by passive stereo-vision-based depth sensors[1]. For inference, we use foreground segmented depth images, which simplifies the proposed generative model (Sect. 4.2.1). The foreground segmentation is obtained by static background subtraction, applied to the raw depth images. For real-world applications with a dynamic background, an online learned background model could be employed (for a comprehensive review of background subtraction methods we refer to [55]). Fig. 4.2 illustrates the intensity images, raw disparity images and foreground segmented depth images, observed from three synchronized sensors.

Typically, the image acquisition time of commodity (depth) cameras can not be synchronized across a network of cameras trivially. To overcome this limitation, we use a pseudo synchronization, where each sensor is synchronized with a global timer over ethernet via the standard Network Time Protocol (NTP). By annotating each captured image with the global acquisition time, it is possible to align the observations from multiple sensors by temporal proximity. As a result we obtain multi-view frames with limited time difference within a single temporal frame. Considering a typical raw frame rate of $25\,\mathrm{Hz}$, the theoretical maximum synchronization

[1] Although our method is particularly able to handle noisy low-resolution depth images, it is not limited to it. Certainly it is also viable to use high-quality depth observations e. g. from an active range sensor.



(a) Sensor view 1 **(b)** Sensor view 2 **(c)** Sensor view 3

(d) Disparity image 1 **(e)** Disparity image 2 **(f)** Disparity image 3

(g) Observation 1 (o_1) **(h)** Observation 2 (o_2) **(i)** Observation 3 (o_3)

Figure 4.2 Example observations from the multi-view setup with five people present in the covered area. Each column corresponds to one sensor. The first row (a-c) shows the gray scale intensity images. In the second row (d-f) the raw disparity images are depicted. The third row (g-i) illustrates the foreground segmented depth images, which are finally used for inference.

error is $\frac{1}{2 \cdot 25}\mathrm{s} = 20\,\mathrm{ms}$. On average the synchronization error in our setup is roughly $10\,\mathrm{ms}$, which is tolerable for the application of people detection. For the rest of this thesis we neglect the synchronization error and assume that the multi-view frames are acquired synchronously.

Even though the models and methods introduced in this thesis are designed to operate on foreground-segmented depth images, they are not limited to these specific observations. By defining a forward model which enables the generation of synthetic observations (cf. Sect. 4.2.1), a different sensor modality can be easily incorporated into the proposed framework. It is even possible to combine different sensor modalities (i.e. types of observations), as long as the forward model for each sensor is well-defined.

4.2 General Model

In this section we define the joint distribution of people present in the scene over space and time. To keep the notation uncluttered, we reuse the symbolic latent variable $\mathcal{X}$ introduced in Sect. 3.1. While in the previous section $\mathcal{X}$ is defined as an arbitrary symbolic latent variable, here we refine the semantics of $\mathcal{X}$. In this section $\mathcal{X}_t$ represents a symbolic scene configuration at time step t. While the abstract scene configuration $\mathcal{X}$ is used to elaborate the general structure of the probabilistic model, it can be instantiated by any kind of scene state representation, including continuous (Sect. 4.3) as well as discrete (Sect. 4.4) representations. The foreground-segmented depth observations at the time step t, acquired from depth sensors $\mathcal{S}_1 \ldots \mathcal{S}_C$, are given as $\mathbf{o}_t = (o_{1,t}, \ldots, o_{C,t})^\intercal$.

The joint probability distribution for time steps $1, \ldots, T$ is given as

$$p(\mathcal{X}_{1:T}, \mathbf{o}_{1:T}) = p(\mathbf{o}_{1:T} \mid \mathcal{X}_{1:T}) p(\mathcal{X}_{1:T}) \,. \tag{4.1}$$

To keep the joint distribution tractable, we make two general assumptions. First, we assume that the probability of the current state $\mathcal{X}_t$ depends only upon on the previous state $\mathcal{X}_{t-1}$ (first order Markov assumption). For $t > 1$ we can write

$$p(\mathcal{X}_t \mid \mathcal{X}_{1:t-1}) = p(\mathcal{X}_t \mid \mathcal{X}_{t-1}) \,. \tag{4.2}$$

Using the Markov assumption, we then can express the distribution of a path of scene configurations as

$$
\begin{aligned}
p(\mathcal{X}_{1:T}) &= p(\mathcal{X}_T \mid \mathcal{X}_{1:T-1}) p(\mathcal{X}_{1:T-1}) && (4.3)\\
&= p(\mathcal{X}_T \mid \mathcal{X}_{T-1}) p(\mathcal{X}_{1:T-1}) && (4.4)\\
&= p(\mathcal{X}_1) \prod_{t=2}^{T} p(\mathcal{X}_t \mid \mathcal{X}_{t-1}) \,. && (4.5)
\end{aligned}
$$

Second, we assume that for given scene configurations $\mathcal{X}_{1:T}$ the observations are conditionally independent

$$
p(\mathbf{o}_{1:T} \mid \mathcal{X}_{1:T}) = \prod_{t=1}^{T} p(\mathbf{o}_t | \mathcal{X}_t) \,. \tag{4.6}
$$

Inserting the immediate results from the first order Markov assumption (4.5) and the conditional independence assumption (4.6) into the joint probability model (4.1), we can write the joint distribution as a first order Hidden Markov Model (HMM)

$$
p(\mathcal{X}_{1:T}, \mathbf{o}_{1:T}) = p(\mathbf{o}_1 \mid \mathcal{X}_1) p(\mathcal{X}_1) \prod_{t=2}^{T} p(\mathbf{o}_t \mid \mathcal{X}_t) p(\mathcal{X}_t \mid \mathcal{X}_{t-1}) \,. \tag{4.7}
$$

Based on the joint distribution (4.7), we will describe the general structure of relevant distributions in the remainder of this section. For ease of notation, we omit the integration domain, where $\int p(\mathcal{X}_{1:T})\, d\mathcal{X}_{1:T}$ is a short hand for $\int_{\mathcal{D}} \cdots \int_{\mathcal{D}} p(\mathcal{X}_{1:T})\, d\mathcal{X}_1 \ldots d\mathcal{X}_T$ with $\mathcal{D}$ defined as the full domain of $\mathcal{X}$. Note that in this section the integrals are defined over an abstract state space implied by $\mathcal{X}$. Depending on the concrete manifestation of the scene configuration space, the integrals have to be refined or, in case of a discrete scene configuration space, replaced by the sum over all discrete states respectively. The distributions derived in the following will find its expression in models for realizations of specific scene configuration spaces in Sect. 4.3 and Sect. 4.4. For a detailed derivation of the general distributions, we refer to Appendix B.

In the context of Bayesian inference, typically the most relevant question is how the latent variables are distributed, given the observations. Formally, conditioning the joint distribution (4.1) on the time series of observations $\mathbf{o}_{1:T}$ leads to the full posterior distribution

$$p(\mathcal{X}_{1:T} \mid \mathbf{o}_{1:T}) = \frac{p(\mathbf{o}_{1:T}, \mathcal{X}_{1:T})}{p(\mathbf{o}_{1:T})} \,. \tag{4.8}$$

In consideration of the first order Markov assumption and the conditional independence of observations (summarized in (4.7)), the posterior can be rearranged for any $t \in \{2, \ldots, T\}$ to

$$p(\mathcal{X}_{1:t} \mid \mathbf{o}_{1:t}) = \frac{p(\mathbf{o}_t \mid \mathcal{X}_t) p(\mathcal{X}_t \mid \mathcal{X}_{t-1}) p(\mathcal{X}_{1:t-1} \mid \mathbf{o}_{1:t-1})}{p(\mathbf{o}_t \mid \mathbf{o}_{1:t-1})} \,, \tag{4.9}$$

as derived in (B.2). In the literature, the denominator $p(\mathbf{o}_t \mid \mathbf{o}_{1:t-1})$ of (4.9) is also referred to as the predicted likelihood and is, according to (B.3a), given as

$$p(\mathbf{o}_t \mid \mathbf{o}_{1:t-1}) = \int p(\mathbf{o}_t \mid \mathcal{X}_t) p(\mathcal{X}_t \mid \mathbf{o}_{1:t-1}) d\mathcal{X}_t \,. \tag{4.10}$$

Since the full posterior distribution given in (4.9) is hard to compute in real world applications due to the integral in the predicted likelihood (4.10), it is often more viable to model the probability of the current state $\mathcal{X}_t$ given the past observations $\mathbf{o}_{1:t}$. Marginalizing the posterior distribution

$p(\mathcal{X}_{1:t} \mid \mathbf{o}_{1:t})$ with respect to the previous states $\mathcal{X}_{1:t-1}$ leads, for any $t \in \{2, \ldots, T\}$, to the recursively defined distribution

$$p(\mathcal{X}_t \mid \mathbf{o}_{1:t}) = \frac{p(\mathbf{o}_t \mid \mathcal{X}_t)}{p(\mathbf{o}_t \mid \mathbf{o}_{1:t-1})} \underbrace{\int p(\mathcal{X}_t \mid \mathcal{X}_{t-1}) p(\mathcal{X}_{t-1} \mid \mathbf{o}_{1:t-1})\, d\mathcal{X}_{t-1}}_{p(\mathcal{X}_t | \mathbf{o}_{1:t-1})}, \tag{4.11}$$

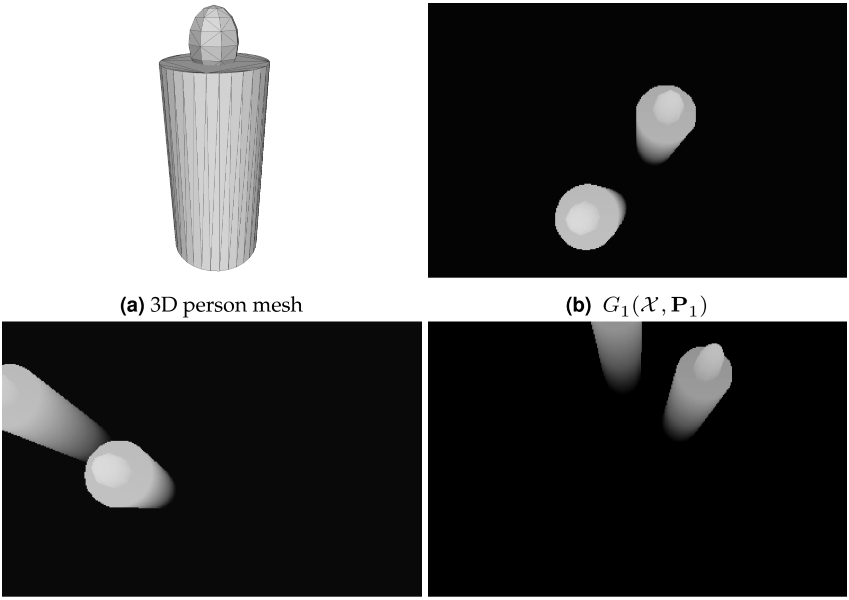
as deduced in (B.4a). In the Bayesian filtering framework this is referred to as the filtering or update distribution, where the integral over the previous state $\mathcal{X}_{t-1}$ in (4.11) is called the predictive distribution

$$p(\mathcal{X}_t \mid \mathbf{o}_{1:t-1}) = \int p(\mathcal{X}_t \mid \mathcal{X}_{t-1}) \underbrace{p(\mathcal{X}_{t-1} \mid \mathbf{o}_{1:t-1})}_{\text{previous filtering dist.}} d\mathcal{X}_{t-1}\,. \tag{4.12}$$

To summarize, rearranging the joint model (4.7) allows to define the recursive filtering distribution $p(\mathcal{X}_t \mid \mathbf{o}_{1:t})$, whereas the previous states are condensed in the predictive distribution $p(\mathcal{X}_t \mid \mathbf{o}_{1:t-1})$. Since the predictive distribution is defined recursively and only includes an integral over the previous state $\mathcal{X}_{t-1}$, it can be approximated effectively in practical applications. For a more general discussion of typical inference problems in Hidden Markov Models we refer to Barber [11, p. 495 ff.].

4.2.1 Generative Scene Model

To define the probabilistic model we make use of a generative scene model $G_c(\mathcal{X}, \mathbf{P}_c)$, which maps a scene configuration $\mathcal{X}$ and a given projection matrix $\mathbf{P}_c$ to a synthetic observation (i.e. synthetic depth image) from the perspective of sensor $\mathcal{S}_c$. We use a simple, rotationally symmetric 3D person model depicted in Fig. 4.3(a), consisting of a cylinder for the body and a sphere for the head. Exemplary samples of the proposed generative scene model are given in Fig. 4.3(b-d).



(a) 3D person mesh

(b) $G_1(\mathcal{X}, \mathbf{P}_1)$

(c) $G_2(\mathcal{X}, \mathbf{P}_2)$

(d) $G_3(\mathcal{X}, \mathbf{P}_3)$

Figure 4.3 (a) 3D polygon mesh used in the generative scene model to represent an individual, where the head is modeled by a sphere and the body by a cylinder. (b-d) Synthetic depth images for a given scene configuration $\mathcal{X}$ from the perspective of each sensor.

Since our generative forward model is not only a function of $\mathcal{X}$, but also of the projection matrix $\mathbf{P}_c$, we incorporate the physical sensor model in a principled way into our framework, allowing us to detect people from arbitrary viewpoints and to easily integrate a new sensor modality.

4.2.2 Data Likelihood

Analog to POM [38] (Sect. 3.2), we assume that the views are conditionally independent with respect to a fixed scene configuration $\mathcal{X}$ to make the likelihood tractable. Since we assume that only people are part of the foreground (cf. Fig. 4.2), and that the depth images are robust regarding illumination changes, this assumption can be justified. Similar to (3.29), the likelihood factorizes as:

$$p(\mathbf{o} \mid \mathcal{X}) = \prod_{c=1}^{C} p(o_c \mid \mathcal{X})\,. \tag{4.13}$$

To model the likelihood $p(o_c \mid \mathcal{X})$ for one depth observation, we assume that an observation can be described by the generative scene model proposed in Sect. 4.2.1 as a functional relation $\mathbf{o}_c = G_c(\mathcal{X}, \mathbf{P}_c)$. Of course, this is a simplification which does not hold in reality for the proposed generative scene model, e. g. due to different shapes and poses of individuals as well as measurement noise. However, it allows us to employ an effective and tractable likelihood model. To weaken this assumption we additionally assume that the observations suffer from measurement noise, thus an observation $\mathbf{o}_c$ is modeled as a realization of a random variable

$$O_c = G_c(\Omega, \mathbf{P}_c) + \eta\,, \tag{4.14}$$

where the random variable η represents the measurement noise and Ω a random variable related to a scene configuration $\mathcal{X}$. In the literature this model is known as *additive noise model* [19, p. 42]. In general, detailed modeling of the depth measurement noise of a stereo vision sensor is challenging because several sources of uncertainty, such as image plane quantization, geometric camera calibration, and stereo correspondences matching, affect the final depth measurement (cf. [16, 51, 72]). For the sake of computational feasibility, we assume additive Gaussian noise[2] , thus

$$\eta \sim \mathcal{N}\Big(0, \sigma_{\text{obs}}^2 \mathbf{I}\Big)\,. \tag{4.15}$$

[2] This implies the assumption that the total measurement noise η is given as a sum of many statistically independent noise sources $\eta = \eta_1 + \cdots + \eta_\nu$. If in addition the Lindeberg condition [57, p. 355 ff.] is satisfied, the Lindeberg-Feller central limit theorem indicates that the distribution of η approaches a normal distribution.

Assuming a fixed scene configuration state $\Omega = \mathcal{X}$, the measurement noise η and the scene configuration Ω are statistically independent $\eta \perp\!\!\!\perp \Omega$ and the only randomness in O_c is the measurement noise η, yielding an observation likelihood

$$\begin{aligned} p(o_c \mid \mathcal{X}, \sigma_{\text{obs}}) &= \mathcal{N}\Big(o_c - G_c(\mathcal{X}, \mathbf{P}_c), \sigma_{\text{obs}}^2 \mathbf{I}\Big) \\ &\propto \exp\left(-\frac{1}{2\sigma_{\text{obs}}^2} \left\| o_c - G_c(\mathcal{X}, \mathbf{P}_c) \right\|_2^2\right). \end{aligned} \tag{4.16}$$

In [19, p. 41 ff.] detailed remarks on the likelihood construction incorporating additive noise can be found.

4.3 Continuous Latent Space

In this section we introduce the first manifestation of the general model presented in the previous Sect. 4.2. The simplified probabilistic model proposed in this section serves two main purposes. On the one hand, it provides an introductory model, for which it is fairly easy to derive the MAP objective and get a point estimate by solving the resulting non-linear least-squares optimization problem (Sect. 5.1.1). On the other hand, the resulting gradient-based inference can be used as a fine-tuning post-processing step in combination with the discrete model and inference methods proposed in Sect. 4.4 and Sec. 5.2, respectively.

For the sake of simplicity, we assume that the number of people m in the scene is known a priori. Once m is fixed, a scene configuration $\mathfrak{X}$ in the continuous latent space can be formally defined as an m-tuple $\mathfrak{X} = (\check{\mathbf{x}}_1, \ldots, \check{\mathbf{x}}_m)$ of ground plane world coordinates $\check{\mathbf{x}}_i \in \mathbb{R}^2$, where each $\check{\mathbf{x}}_i$ corresponds to the location of an individual. Based on the general joint model (4.7) and the independence assumption (4.13), we define the posterior distribution

$$p(\mathcal{X}_{1:T} \mid \mathbf{o}_{1:T}) = \frac{\prod_{t=1}^{T} \left[\prod_{c=1}^{C} p\big(o_{c,t} \mid \mathcal{X}_t\big)\right] p(\mathcal{X}_t \mid \mathcal{X}_{t-1}) p(\mathcal{X}_t)}{p(\mathbf{o}_{1:T})}, \quad (4.17)$$

with the definition $p(\mathcal{X}_1 \mid \mathcal{X}_0) \overset{\text{def.}}{=} 1$. In comparison to the general joint distribution (4.7), we append a prior distribution $p(\mathcal{X}_t)$ to (4.17), enabling the incorporation of a-priori scene knowledge. The data likelihood at one time step $\prod_{c=1}^{C} p\big(o_{c,t} \mid \mathcal{X}_t\big)$ is identical to the general model as introduced in Sect. 4.2.2. For inference in continuous latent space we are only interested in the MAP state, therefore we refrain from further consideration of the evidence $p(\mathbf{o}_{1:T})$ in this section. The two remaining terms, namely the dynamics model $p(\mathcal{X}_t \mid \mathcal{X}_{t-1})$ and the prior $p(\mathcal{X}_t)$, are defined in the following.

4.3.1 Dynamics Model

We employ a dynamics model to express the relation between two consecutive scene configurations. It is assumed that the movements of individuals are independent of each other, thus the dynamics term factorizes over the individual person locations $\check{\mathbf{x}}_{i,t}$,

$$p(\mathcal{X}_t \mid \mathcal{X}_{t-1}) = \prod_{i=1}^{m} p\big(\check{\mathbf{x}}_{i,t} \mid \check{\mathbf{x}}_{i,t-1}\big). \quad (4.18)$$

Since we assume only small movements of persons between two consecutive frames, we do not consider the velocity of individuals, but employ a first-order motion model. Hence, it is assumed that a person location at time step $\check{\mathbf{x}}_{i,t}$ is a noisy version of its predecessor $\check{\mathbf{x}}_{i,t-1}$. To further simplify the inference on this model, we assume that $\check{\mathbf{x}}_{i,t}$ is normal distributed around $\check{\mathbf{x}}_{i,t-1}$. Thus the dynamics distribution for an individual person location $\check{\mathbf{x}}_{i,t}$ is given as

$$p\big(\check{\mathbf{x}}_{i,t} \,\big|\, \check{\mathbf{x}}_{i,t-1}, \boldsymbol{\Sigma}\big) = \mathcal{N}\big(\check{\mathbf{x}}_{i,t} \,\big|\, \check{\mathbf{x}}_{i,t-1}, \boldsymbol{\Sigma}\big)\,. \tag{4.19}$$

In Sect. 5.1.1 it turns out that this simplified Gaussian dynamics model leads to an effective and tractable objective for MAP inference. For a broad discussion of motion models in the context of people tracking we refer to Fleet [37].

4.3.2 A Priori Assumptions

To incorporate given knowledge of the underlying scenario, we add prior terms to further restrict the set of likely scene states. For a scene configuration at one time step $\mathcal{X}_t$ we employ two independent prior assumptions, $p_{\text{box}}(\mathcal{X}_t)$ and $p_{\text{dist}}(\mathcal{X}_t)$, with $p(\mathcal{X}_t) = p_{\text{box}}(\mathcal{X}_t)p_{\text{dist}}(\mathcal{X}_t)$. Since the prior terms are independent of the current time step, we omit the time index t in this section.

The first prior term $p_{\text{box}}(\mathcal{X})$ reflects our knowledge of the visible ground plane (which can be inferred by the sensor network calibration, cf. 4.1.1). We model this knowledge by assuming that a person location $\check{\mathbf{x}}_i$ is uniformly distributed in the observable rectangular area,

$$p_{\text{box}}(\mathcal{X}) = \prod_{i=1}^{m} p(\check{\mathbf{x}}_i) \tag{4.20}$$

with

$$p(\check{\mathbf{x}}_i) = \mathcal{U}(\check{\mathbf{x}}_{\text{min}}, \check{\mathbf{x}}_{\text{max}})\,, \tag{4.21}$$

where $\mathcal{U}(\check{\mathbf{x}}_{\text{min}}, \check{\mathbf{x}}_{\text{max}})$ is the uniform distribution over the approximated rectangular area. In consequence the probability $p(\check{\mathbf{x}}_i)$ is zero for $\check{\mathbf{x}}_i$ being outside of the observable area.

The second prior can be motivated by the interpersonal distance. Depending on the social context, two individuals in general keep a certain distance to each other in order to preserve their private space. We therefore consider the distance between all possible pairs of individuals in the scene. Formally, the distance prior factorizes over all possible pairs of person locations

$$p_{\text{dist}}(\check{\mathbf{x}}_1, \ldots, \check{\mathbf{x}}_m) = \prod_{i=1}^{m-1} \prod_{j=i+1}^{m} p(\check{\mathbf{x}}_i, \check{\mathbf{x}}_j) \,. \tag{4.22}$$

Modeling the joint probability between two locations as a zero mean normal distribution with respect to the inverse distance

$$d(\check{\mathbf{x}}_i, \check{\mathbf{x}}_j) = \frac{1}{\left\|\check{\mathbf{x}}_i - \check{\mathbf{x}}_j\right\|_2 + \epsilon} \,, \tag{4.23}$$

we can write the pairwise joint probability as the unnormalized pseudo distribution

$$\tilde{p}\big(\check{\mathbf{x}}_i, \check{\mathbf{x}}_j \bigm| \sigma_{\text{dist}}\big) \propto \exp\!\left(-\frac{1}{2\sigma_{\text{dist}}^2} \left\|d(\check{\mathbf{x}}_i, \check{\mathbf{x}}_j)\right\|_2^2\right) , \tag{4.24}$$

with σ_{dist} being the std. deviation of the inverse distance. To calculate the partition function this distribution needs to be defined on a finite interval. However, this is not relevant in this case, since for this model we are only interested in an MAP point estimate, which does not depend on the constant normalization term anyway. In Fig. 4.4 an exemplary function plot of the unnormalized distribution (4.24) is depicted for a fixed person location $\check{\mathbf{x}}_j$. It is clearly recognizable that the proposed pseudo distribution has the desired effect: the distance between two individuals $\check{\mathbf{x}}_i, \check{\mathbf{x}}_j$ is positive correlated to the pairwise probability density $\tilde{p}\big(\check{\mathbf{x}}_i, \check{\mathbf{x}}_j \bigm| \sigma_{\text{dist}}\big)$.

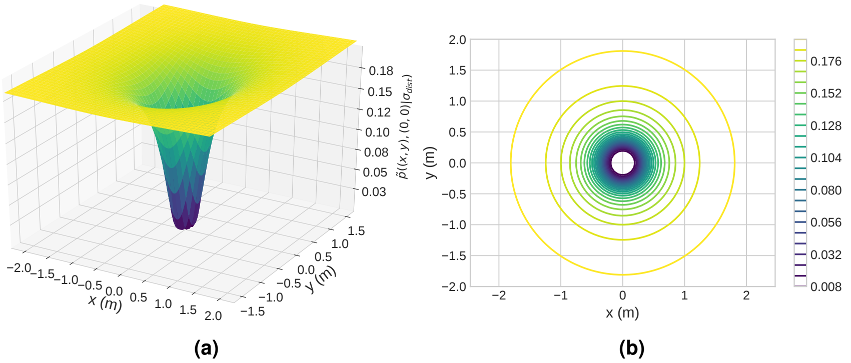


Figure 4.4 Exemplary surface plot (a) and contour plot (b) for the pairwise distance prior. The unnormalized density $\tilde{p}(\check{\mathbf{x}}_i = (x, y), \check{\mathbf{x}}_j = (0, 0) \mid \sigma_{\text{dist}} = 2)$ is plotted as a function of ground plane coordinates (x, y) with respect to a fixed individual at $\check{\mathbf{x}}_j = (0, 0)$.

4.4 Discrete Latent Space

The continuous latent space, as introduced in the previous section, has strong limitations. First, the number of people in the scene has to be known a priori. Second, in Sect. 5.1 it turns out that the final MAP inference highly depends on a good initialization due to the underlying gradient based continuous optimization. To overcome these mentioned shortcomings, we propose a discrete scene configuration space. Similar to Fleuret *et al.* [38], the basic idea is to model the ground level plane as a discrete grid, where each grid cell can either be occupied by a person or be empty. In the robotic community, the general concept of a discrete grid of binary occupancy states is also referred to as an *occupancy map* (cf. Thrun [94]). In contrast to the continuous latent space model, and as a direct consequence of the occupancy map, there is no need to explicitly model the number of individuals in the scene to define a latent space of fixed dimensions. Moreover, the discrete scene configuration space allows us to define a probabilistic model in which approximate inference can be obtained effectively by mean-field variational inference (cf. Sect. 5.2). For the formal definition of the discrete probabilistic model, the ground plane area is discretized

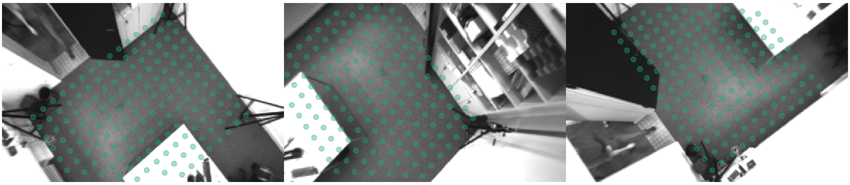


Figure 4.5 Discrete ground plane grid, projected into all three sensor views. Each grid location (turquoise point) is associated with one Bernoulli random variable $X_{i,t}$, reflecting if a person is occupying the grid cell at time step t.

into a 2D-grid of n locations (see Fig. 4.5). Each grid location u_i will be assigned a realization $x_{i,t}$ of a Bernoulli random variable $X_{i,t} \sim \mathcal{B}(\mu_{i,t})$, where $\mu_{i,t}$ denotes the probability of a person present at location u_i at time step t with $1 \leq t \leq T$. The scene configuration for one time step t is then given as the vector $\mathbf{x}_t = (x_{1,t}, \ldots, x_{n,t})^\mathsf{T} \in \{0,1\}^n$ (cf. POM [38], Sect. 3.2.1).

Following the assumptions of the general probabilistic model defined in Sect. 4.2, the joint distribution for a sequence of discrete scene configurations $\mathbf{x}_{1:T}$ is given as

$$p(\mathbf{x}_{1:T}, \mathbf{o}_{1:T}) = \prod_{t=1}^{T} \prod_{c=1}^{C} p\big(o_{c,t} \mid \mathbf{x}_t\big) p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) \,, \tag{4.25}$$

with $p(\mathbf{x}_1 \mid \mathbf{x}_0) \overset{\text{def.}}{=} 1$. As a direct consequence of the joint distribution (4.25) and the general distributions in (4.9)–(4.11), we formulate three distributions for the discrete model: (i) the data posterior distribution, which omits the temporal context; (ii) the full posterior distribution; and (iii) the recursively defined Bayesian filtering distribution. For all three distributions we propose approximate inference methods in Sect. 5.2.

(i) Data posterior $p(\mathbf{x}_t \mid \mathbf{o}_t)$ Omitting the temporal relationship between consecutive scene states, the data posterior distribution allows us to express the likeliness of a scene configuration x_t, given the observations from all sensors at one time step $\mathbf{o}_t$. Assuming that the prior for a scene configuration factorizes as $p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i)$, the posterior distribution is defined as

$$p(\mathbf{x}_t \mid \mathbf{o}_t) = \frac{\prod_{c=1}^{C} p(o_{c,t} \mid \mathbf{x}_t) \prod_{i=1}^{n} p(x_{i,t})}{\sum_{\mathbf{x}' \in \{0,1\}^n} \prod_{c=1}^{C} p(o_{c,t} \mid \mathbf{x}') \prod_{i=1}^{n} p(x_i')} . \tag{4.26}$$

The data likelihood follows directly from the general model defined in Sect. 4.2.2.

(ii) Full posterior $p(\mathbf{x}_{1:T} \mid \mathbf{o}_{1:T})$ The full posterior takes all observations from time steps $1, \ldots, T$ into account and models the likeliness of a path of scene configurations $\mathbf{x}_1, \ldots, \mathbf{x}_T$. The full posterior is given as

$$p(\mathbf{x}_{1:T} \mid \mathbf{o}_{1:T}) = \frac{\prod_{t=1}^{T} \prod_{c=1}^{C} p(o_{c,t} \mid \mathbf{x}_t) p(\mathbf{x}_t \mid \mathbf{x}_{t-1})}{p(\mathbf{o}_{1:T})} , \tag{4.27}$$

with $p(\mathbf{x}_1 \mid \mathbf{x}_0) \overset{\text{def.}}{=} 1$. In addition to the data posterior, the full posterior includes a state transition distribution $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ which represents the evolution of scene configuration states over time. The underlying probabilistic dynamics model is defined in Sect. 4.4.1. Note that the full posterior is hard to compute in real-world scenarios since the complexity grows with the number of time steps. However, during inference it can serve as a baseline for comparison with the recursively defined filtering distribution and can find its application in scenarios where offline batch processing[3] is viable.

(iii) Bayesian filtering distribution $p(\mathbf{x}_t \mid \mathbf{o}_{1:t})$ In contrast to the full posterior, the filtering distribution models the likeliness of the current scene configuration state $\mathbf{x}_t$, given all observations $\mathbf{o}_{1:t}$ up to time t. Applying the Bayesian filtering framework allows the recursive definition of the filtering distribution, where the past is condensed

[3] In this context batch processing refers to a detection method taking a sequence of frames as input rather than operating in a typical frame-by-frame manner.

in the filtering distribution $p(\mathbf{x}_{t-1} \mid \mathbf{o}_{1:t-1})$ from the previous time step (as described in Sect. 4.2). Applying (4.11) from the general model, the filtering distribution for the discrete scene configuration space is given as

$$p(\mathbf{x}_t \mid \mathbf{o}_{1:t}) = \frac{p(\mathbf{o}_t \mid \mathbf{x}_t) p(\mathbf{x}_t \mid \mathbf{o}_{1:t-1})}{p(\mathbf{o}_t \mid \mathbf{o}_{1:t-1})} \, , \tag{4.28}$$

with the predictive distribution

$$p(\mathbf{x}_t \mid \mathbf{o}_{1:t-1}) = \sum_{\mathbf{x}_{t-1} \in \{0,1\}^n} p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) \cdot \underbrace{p(\mathbf{x}_{t-1} \mid \mathbf{o}_{1:t-1})}_{\text{previous filtering dist.}} \, . \tag{4.29}$$

4.4.1 Grid Dynamics Model

The complete construction of the distributions over space and time (4.27), (4.28), requires the definition of a state transition model $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ that describes the evolution of scene configurations over time. In this work we do not focus on tracking but on leveraging the temporal context to regularize the mean-field optimization (see Sect. 5.2). Therefore, we propose a grid based dynamics model, without modeling the explicit motion of objects. This leads to a computationally convenient model which represents the flow of probability mass across space and time.

Computationally feasible inference can be achieved by assuming conditional independence of individual grid cell states $x_{t,i}$, given the previous state $\mathbf{x}_{t-1}$. Hence, the state transition distribution factorizes as

$$p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \prod_{j=1}^{n} p\big(x_{j,t} \bigm| \mathbf{x}_{t-1}\big) \, . \tag{4.30}$$

This assumption is limiting the expressiveness of our model significantly because it prevents from modeling the relationship between grid cells at the current time step t. This can be illustrated by the following example.

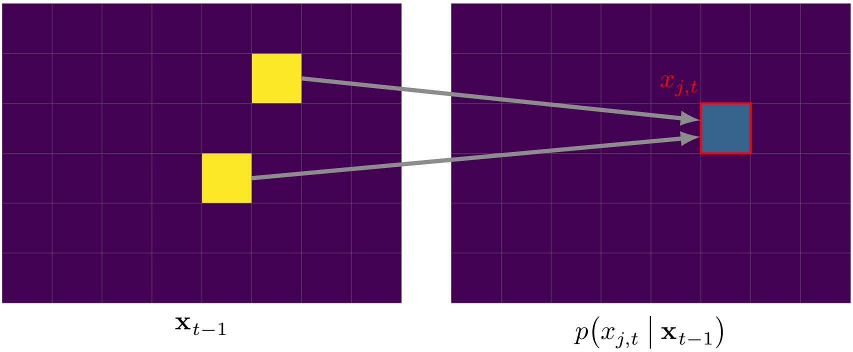


Figure 4.6 Schematic illustration of the transition distribution $p(x_{j,t} \mid \mathbf{x}_{t-1})$. For an exemplary grid cell u_j the probability $p(x_{j,t} \mid \mathbf{x}_{t-1})$ (right image, red marked cell) is defined as a weighted sum over the transitions from all grid cells occupied in the previous time step (left image, yellow grid cells).

Consider a person present at grid location u_k at time step $t-1$, moving to grid location u_i or u_j in the next time step t. Regarding the factorization (4.30) it is only possible to assign two independent probabilities, $p(x_{i,t} \mid \mathbf{x}_{t-1})$ for the person present at grid location u_i and $p(x_{j,t} \mid \mathbf{x}_{t-1})$ for the person present at grid location u_j, respectively. However, it is not possible to express that the person moves to either grid location u_i or u_j but cannot occupy both grid cells at time step t. To be able to model the joint probability $p(x_{i,t}, x_{j,t} \mid \mathbf{x}_{t-1})$ at least a pairwise factorization of (4.30) is required, thus $p(\mathbf{x} \mid \mathbf{x}_{t-1}) = \prod_{j \sim i} p(x_{i,t}, x_{j,t} \mid \mathbf{x}_{t-1})$, with $j \sim i$ denoting all pairwise direct neighbor indices. However, such an assumption would strongly increase the complexity of our model, leading to a computationally intensive inference problem, despite the mean-field approximation introduced in Sect. 5.2. Therefore, we leave this as an open question for future work and stick to the simplification (4.30) for this work.

Considering the transition distribution (4.30) we need to define the probability $p(x_{j,t} \mid \mathbf{x}_{t-1})$ of a person present at a grid cell u_j at time step t, given the previous scene configuration state $\mathbf{x}_{t-1}$. As illustrated in

Fig. 4.6, the basic idea is to express the distribution $p\big(x_{j,t} \mid \mathbf{x}_{t-1}\big)$ as a weighted sum of the transitions from all previous $x_{i,t-1}$ being in state one (meaning that a person is present). The desired probability distribution is formally derived by first adding a hidden latent auxiliary variable to the joint model and then marginalizing over this variable[4]. More precisely, we first introduce an auxiliary latent random variable $\mathbf{Z}$ to the distribution $p\big(x_{j,t} \mid \mathbf{x}_{t-1}\big)$, indicating the presence of a single individual in the previous state. Second, we marginalize over $\mathbf{Z}$ to express $p\big(x_{j,t} \mid \mathbf{x}_{t-1}\big)$ as a sum over all transitions from grid cells occupied in the previous state $\mathbf{x}_{t-1}$.

Let $\mathbf{Z}$ be a one-hot-encoded random variable with the realizations being $\mathbf{z} = (z_1, \dots, z_n)^\intercal$ with $z_k \in \{0, 1\}$, such that $\sum_{k=1}^{n} z_k = 1$. Further, let $p(z_k = 1)$ refer to the probability of a realization $\mathbf{z}$ with exactly one non-zero element z_k. As a consequence of the one-hot encoding the distribution of $\mathbf{z}$ can be written as

$$p(\mathbf{z}) = \prod_{k=1}^{n} p(z_k)^{z_k} \,, \tag{4.31}$$

with the definition $0^0 \overset{\text{def.}}{=} 1$. For the ease of notation we introduce weighting coefficients $w_k \in [0, 1]$ with $\sum_{k=1}^{n} w_k = 1$ such that $p(z_k = 1) = w_k$, thus (4.31) can be rewritten as

$$p(\mathbf{z}) = \prod_{k=1}^{n} w_k^{z_k} \,. \tag{4.32}$$

Introducing $\mathbf{z}$ to the distribution $p\big(x_{j,t} \mid \mathbf{x}_{t-1}\big)$ leads to the joint distribution

$$p\big(x_{j,t}, \mathbf{z} \mid \mathbf{x}_{t-1}\big) = p\big(x_{j,t} \mid \mathbf{z}, \mathbf{x}_{t-1}\big) p(\mathbf{z} \mid \mathbf{x}_{t-1}) \,. \tag{4.33}$$

[4] This principle is well know in probabilistic modeling, e. g. a similar technique is used for mixture of experts models (cf. Barber [11, p. 448 ff.], Murphy [70, p. 344 ff.]).

Semantically z_k is equal to one if a person is present at cell u_k in the previous scene configuration $\mathbf{x}_{t-1}$, thus $x_{k,t-1} = 1$. In consideration of the one-hot encoding of z and (4.31) it follows

$$p\big(x_{j,t} \bigm| \mathbf{z}, \mathbf{x}_{t-1}\big) = \prod_{k=1}^{n} p\big(x_{j,t} \bigm| z_k = 1, \cancel{\mathbf{x}_{t-1}}\big)^{z_k} \tag{4.34}$$

$$= p\big(x_{j,t} \bigm| z_k = 1\big)\,, \tag{4.35}$$

where we omit the conditioning on $\mathbf{x}_{t-1}$. The resulting discrete distribution $p\big(x_{j,t} \bigm| z_k = 1\big)$ reflects the probability of $x_{j,t}$ given that one particular cell with index k is one in the previous state $\mathbf{x}_{t-1}$. Marginalization of (4.33) with respect to $\mathbf{z}$ results in the mixture model

$$p\big(x_{j,t} \bigm| \mathbf{x}_{t-1}\big) = \sum_{\mathbf{z} \in \{0,1\}^n : |\mathbf{z}|=1} p\big(x_{j,t}, \mathbf{z} \bigm| \mathbf{x}_{t-1}\big) \tag{4.36}$$

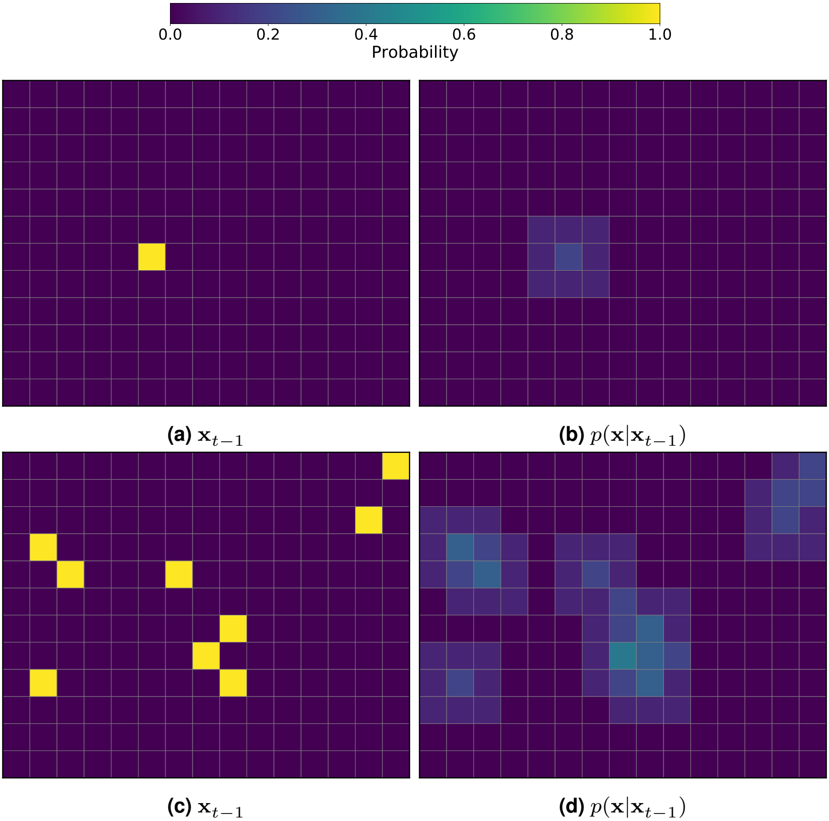
$$= \sum_{k=1}^{n} p\big(x_{j,t} \bigm| z_k = 1\big) p(z_k = 1 \mid \mathbf{x}_{t-1}) \tag{4.37}$$

$$= \sum_{k=1}^{n} p\big(x_{j,t} \bigm| z_k = 1\big) \cdot w_k\,. \tag{4.38}$$

According to the definition of the weighting coefficients w_k (cf. (4.32)), the distribution $p(z_k = 1 \mid \mathbf{x}_{t-1})$ can be interpreted as normalization weights

$$w_k = \begin{cases} \frac{1}{\|\mathbf{x}_{t-1}\|_1}, & \text{if } x_{k,t-1} = 1 \\ 0, & \text{else }. \end{cases} \tag{4.39}$$

Finally, the probability flow depends on the definition of the transition distribution $p\big(x_{j,t} \bigm| z_k = 1\big)$ in (4.38), which denotes the probability that a person is present at location u_j, given that a person was present at location u_k at the previous time step. In our setup we expect only limited movement of individuals between two consecutive frames. Hence, we assume



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

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

(c) $\mathbf{x}_{t-1}$

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

Figure 4.7 Example of proposed discrete dynamics model for two scene configurations (a,c) with the corresponding output distributions (b,d) for $\mu_{\text{self}} = 0.2$ and $w_k = 1$.

individuals will either move to an adjacent grid cell with a probability of μ_{ne}, or stay at the current cell with a probability of μ_{self} (see Fig. 4.7). Let

$\mathcal{B}(\cdot \mid \mu)$ be the probability mass function of a Bernoulli distribution with the parameter μ, then the transition distribution is given as

$$p\big(x_{j,t} \mid z_k = 1\big) = \begin{cases} \mathcal{B}\big(x_{j,t} \mid \mu_{\text{self}}\big) \ , & \text{if } j = k \\ \mathcal{B}\big(x_{j,t} \mid \mu_{\text{ne}}\big) \ , & \text{if } j \in N_k \\ 1 - x_{j,t} \ , & \text{else} \ , \end{cases} \tag{4.40}$$

with N_k being the index set of the direct neighbors of u_k. We define

$$\mu_{\text{ne}} = \frac{1 - \mu_{\text{self}}}{|N_k|} \ , \tag{4.41}$$

which leads to the special case where the emitted probability for one person present at the previous time step equals to one. As a direct consequence of the chosen transition model (4.40), which only allows movement in the direct neighborhood, we can set

$$w_k = \begin{cases} 1, & \text{if } x_{k,t-1} = 1 \\ 0, & \text{else} \ , \end{cases} \tag{4.42}$$

while (4.38) still meets the requirements of a probability mass function. This has the additional effect that the expected number of people in the scene with respect to the dynamics model stays constant, thus

$$\langle \|\mathbf{x}_t\|_1 \rangle_{p\left(\mathbf{x} \mid \mathbf{x}_{t-1}\right)} = \|\mathbf{x}_{t-1}\|_1 \ . \tag{4.43}$$

Note that because of the normalization weights w_k it is generally also possible to use more sophisticated transition probability distributions. In Fig. 4.7 two concrete samples of the proposed dynamics distribution $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ are given. Fig. 4.7(a),(b) show the trivial case with only one person present at cell u_j in the previous time step $t-1$. According to (4.40) the probability mass is distributed uniformly with $\mu_{\text{ne}} = 0.1$ to the direct eight neighbors and is 0.2 at cell u_j at time step t. In Fig. 4.7(c),(d) a more complex example with $\|\mathbf{x}_{t-1}\|_1 = 9$ persons present at the previous time step $t-1$ is presented.

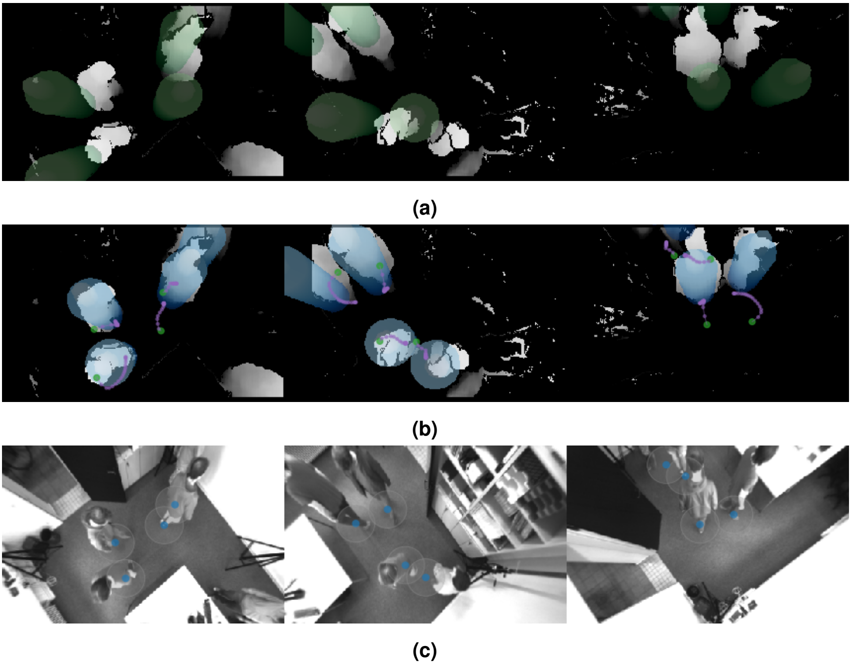
5 Inference

In the previous chapter we proposed different probabilistic models for people detection in multiple depth images, introducing assumptions about the scene and the sensor observations. In this chapter we will focus on the probabilistic inference[1] regarding those models. In Sect. 5.1 we derive the MAP objective for the continuous model introduced in Sect. 4.3 and show how the final non-linear least squares problem can be solved by gradient based optimization methods. Instead of just obtaining an MAP point estimate, in Sect. 5.2 we propose to use mean-field variational inference to approximate the varieties of discrete probability distributions introduced in Sect. 4.4. In contrast to the aforementioned generative probabilistic inference methods, in Sect. 5.3 we propose a discriminatively trained multi-view CNN architecture, allowing a direct comparison between generative and discriminative inference in Chapter 6.

5.1 Maximum a Posterior Inference in Continuous Latent Space

In this section we present an inference method for the continuous scene model introduced in Sect. 4.3. For discrete scene configuration space, the proposed inference method allows the approximation of the full posterior distribution. In contrast, for the continuous case we only obtain a maximum a posterior (MAP) point estimate. As defined in Sect. 4.3, it is

[1] Probabilistic inference refers to the task of estimating the probability distribution (or in case of MAP inference just the mode of the distribution) of one or more latent variables, given some evidence (observed variables).



(a)

(b)

(c)

Figure 5.1 Optimization result for MAP objective without temporal dynamics. (a) shows the depth observations with the synthetic depth images corresponding to the initial scene configuration as green overlay. (b) shows the synthetic depth images corresponding to the final MAP result as blue overlay. The initial positions are given as green dots, the intermediate states of the optimization are drawn in purple. (c) illustrates the final optimization result back projected into the camera image of each sensor.

assumed that the number of people in the scene is known a priori for the continuous model. This assumption is crucial for the proposed inference method, since it allows straight forward gradient based optimization on the location estimates of individuals in the scene. As a consequence of this, the proposed method is not suitable for inference in real-world applications without further pre-processing (e. g. estimating the number of people in the scene by a coarse person detection method). Still, it can be seen as a complementary method to the proposed discrete inference method, since it effectively enables fine-tuned discrete person localization (cf. Sect. 6.6, in particular Fig. 6.20). In the following we first derive the MAP objective (Sect. 5.1.1) and subsequently focus on practically solving the resulting non-linear least squares optimization problem (Sect. 5.1.2). In Fig. 5.1 an exemplary optimization process for one multi-view frame is depicted.

5.1.1 MAP Objective

With respect to the continuous probabilistic model defined in (4.17), the maximum a posteriori (MAP) objective is to estimate a sequence of scene configurations $\hat{\mathfrak{X}}_{1:T}$, which most likely explains the sequence of observations $\mathbf{o}_{1:T}$. According to (4.17) the posterior distribution can be restated as

$$p(\mathfrak{X}_{1:T} \mid \mathbf{o}_{1:T}) \propto \prod_{t=1}^{T} \underbrace{\left[\prod_{c=1}^{C} p\big(o_{c,t} \mid \mathfrak{X}_t\big)\right]}_{\text{data likelihood}} p(\mathfrak{X}_t \mid \mathfrak{X}_{t-1}) p(\mathfrak{X}_t)\,. \tag{5.1}$$

The mode of $p(\mathfrak{X}_{1:T} \mid \mathbf{o}_{1:T})$ is referred to as the maximum a posterior scene configuration. In consequence the MAP objective is formally given as

$$\hat{\mathfrak{X}}_{1:T} = \underset{\mathfrak{X}_{1:T}}{\arg\max}\; p(\mathfrak{X}_{1:T} \mid \mathbf{o}_{1:T}) \tag{5.2}$$

$$= \underset{\mathfrak{X}_{1:T}}{\arg\max}\; \log\left(\prod_{t=1}^{T} \left[\prod_{c=1}^{C} p\big(o_{c,t} \mid \mathfrak{X}_t\big)\right] p(\mathfrak{X}_t \mid \mathfrak{X}_{t-1}) p(\mathfrak{X}_t)\right), \tag{5.3}$$

where in (5.3) we substituted the logarithm of (5.1). Applying the product rule of logarithms (5.4) and inserting the data likelihood and the dynamics model (5.5), we can recast the objective as a non-linear least squares optimization problem

$$\begin{aligned}\hat{\mathfrak{X}}_{1:T} &= \underset{\mathfrak{X}_{1:T}}{\arg\max} \sum_{t=1}^{T}\sum_{c=1}^{C} \log\big(p\big(o_{c,t} \mid \mathfrak{X}_t\big)\big) + \sum_{t=1}^{T} \log(p(\mathfrak{X}_t \mid \mathfrak{X}_{t-1})) \\ &\quad + \sum_{t=1}^{T} \log(p(\mathfrak{X}_t)) \qquad (5.4) \\ &= \underset{\mathfrak{X}_{1:T}}{\arg\min} \sum_{t=1}^{T}\sum_{c=1}^{C} \|o_c - G_c(\mathfrak{X}_t, \mathbf{P}_c)\|_2^2 + \sum_{t=2}^{T} \|\mathfrak{X}_t - \mathfrak{X}_{t-1}\|_2^2 \\ &\quad \underbrace{+ \sum_{t=1}^{T} E_t^{\text{box}} + \sum_{t=1}^{T} E_t^{\text{dist}}}_{\text{regularization}}, \qquad (5.5)\end{aligned}$$

where the terms $E_{\text{box}}, E_{\text{dist}}$ in (5.5) correspond to the box prior and the distance prior defined in (4.20) and (4.22) respectively. Since the likelihood and prior terms defined in (4.13) and (4.20, 4.22) only depend on the scene configuration $\mathfrak{X}_t$ at a single time step, they are aggregated as a sum over each time step t.

From an optimization perspective, the prior terms act as regularization terms, restricting the set of possible scene configurations. The term E_{box} penalizes person locations $\check{\mathbf{x}}$ outside of the observable area. Continuous numerical optimization methods require differentiable energy terms. Therefore, we approximate the box penalty by a function which is zero for person locations in the observable area and increases with the distance to the border of the observable area. Let $\check{\mathbf{x}}_i^\nu$ be ν-th component of $\check{\mathbf{x}}_{i,t}$, then the box penalty is given as

$$E_t^{\text{box}} = \sum_{i=1}^{m} \sum_{\nu \in \{1,2\}} \left[\max(\check{\mathbf{x}}_{\text{min}}^\nu - \check{\mathbf{x}}_i^\nu, 0) + \max(\check{\mathbf{x}}_i^\nu - \check{\mathbf{x}}_{\text{max}}^\nu, 0)\right]^2, \qquad (5.6)$$

where $\check{\mathbf{x}}_{min}, \check{\mathbf{x}}_{max}$ denote the borders of the observable area.

The distance energy term is a direct result of the prior proposed in (4.22), with an additional $\max$ function, which assigns costs to all pairwise distances smaller than $\kappa = 1\,\mathrm{m}$:

$$E_t^{\text{dist}} = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \left[\max\left(d(\check{\mathbf{x}}_{i,t}, \check{\mathbf{x}}_{j,t}) - \frac{1}{\kappa}, 0 \right) \right]^2 . \tag{5.7}$$

We introduce the weighting parameters $\lambda_{\text{temporal}}, \lambda_{\text{box}}, \lambda_{\text{dist}} \in \mathbb{R}^+$ to balance the impact of the individual energy terms. The final non-linear least-squares objective for estimating the MAP path of scene configurations $\mathfrak{X}_{1:T}$ given a sequence of observations $\mathbf{o}_{1:T}$ (cf. posterior distribution (4.17)) is given as:

$$\begin{aligned} \hat{\mathfrak{X}}_{1:T} = \underset{\mathfrak{X}_{1:T}}{\arg\min} & \underbrace{\sum_{t=1}^{T} \sum_{c=1}^{C} \| o_c - G_c(\mathfrak{X}_t, \mathbf{P}_c) \|_2^2}_{\text{data likelihood}} + \underbrace{\lambda_{\text{temporal}} \sum_{t=2}^{T} \| \mathfrak{X}_t - \mathfrak{X}_{t-1} \|_2^2}_{\text{temporal dynamics}} \\ & \underbrace{+ \lambda_{\text{box}} \sum_{t=1}^{T} E_t^{\text{box}} + \lambda_{\text{dist}} \sum_{t=1}^{T} E_t^{\text{dist}}}_{\text{a priori assumptions}} . \end{aligned} \tag{5.8}$$

The temporal dynamics term connects a scene configuration $\mathfrak{X}_t$ with the preceding scene configuration $\mathfrak{X}_{t-1}$, enabling joint optimization of the scene configurations $\mathfrak{X}_{1:T}$. The dynamics term can be interpreted as a temporal smoothing term, ensuring temporally consistent scene configurations. In fact, qualitative experiments (Sect. 6.6.2) show that this smoothing property can effectively prevent the optimization from getting stuck in a local minimum.

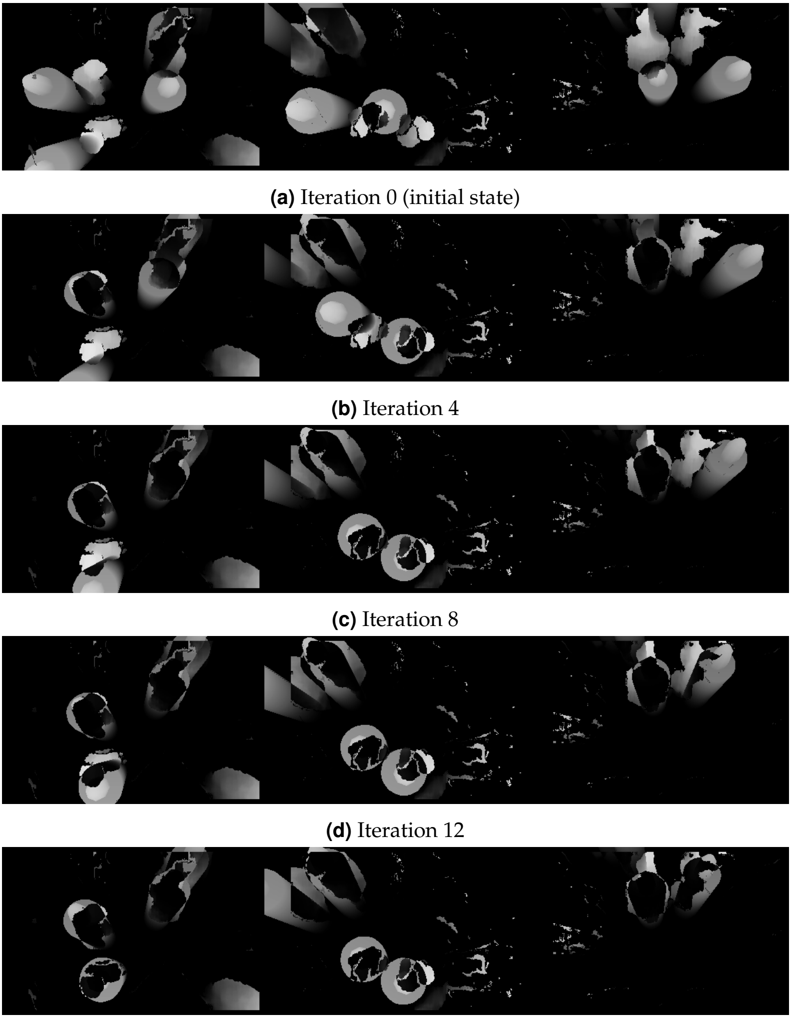
For single frame inference the temporal term can be dropped and the objective for a scene configuration at a single time step trivially follows from the full objective (5.8):

$$\begin{aligned}\hat{\mathfrak{X}}_t &= \arg\max_{\mathfrak{X}_t} p(\mathfrak{X}_t \mid \mathbf{o}_t) \\ &= \arg\min_{\mathfrak{X}_t} \sum_{c=1}^{C} \|o_c - G_c(\mathfrak{X}_t, \mathbf{P}_c)\|_2^2 + \lambda_{\text{box}} E_t^{\text{box}} + \lambda_{\text{dist}} E_t^{\text{dist}} \,.\end{aligned} \tag{5.9}$$

5.1.2 Optimization

Estimating the MAP sequence of scene configurations is obtained by numerically solving the non-linear least squares (NLLSQ) objective (5.8), using a continuous optimization method. In this work we use the iterative NNLSQ solver *dogleg* [76, 77]. Dogleg is a trust-region method, combining the Gauss-Newton algorithm with gradient descent. Notice that our objective does not depend on a specific solver, thus one could also use other NLLSQ solving methods, such as the well-known Levenberg–Marquardt algorithm. For a comprehensive discussion on NLLSQ solving methods we refer to [71, p. 254 ff.].

Of-the-shelf gradient based NLLSQ methods require that each part of the objective has to be differentiable. Regarding the objective (5.8) this is the case for the regularization terms. The crucial part of (5.8) is the data likelihood term $\|o_c - G_c(\mathfrak{X}_t, \mathbf{P}_c)\|_2^2$ (cf. residual images in Fig. 5.2). Calculating the gradients for every output pixel with respect to $\mathfrak{X}_t$ requires differentiating the generative scene model function $G_c(\mathfrak{X}_t, \mathbf{P}_c)$, which maps a scene configuration $\mathfrak{X}_t$ to its corresponding synthetic depth image (cf. Sect. 4.2.1). More precisely this involves the calculation of the partial derivatives for every output pixel $G_{c,j}(\mathfrak{X}_t, \mathbf{P}_c)$ with respect to each scene configuration variable $\mathfrak{X}_t = (\check{\mathbf{x}}_1, \ldots, \check{\mathbf{x}}_m)^\intercal = (\check{x}_{1,1}, \check{x}_{1,2}, \ldots, \check{x}_{m,1}, \check{x}_{m,2})^\intercal \in$



(a) Iteration 0 (initial state)

(b) Iteration 4

(c) Iteration 8

(d) Iteration 12

(e) Iteration 16

Figure 5.2 Residual images showing the optimization process for a single time step. Each column corresponds to one sensor c and each row to an optimization iteration. The images show the normalized per-pixel error $\|o_c - G_c(\mathcal{X}_t, \mathbf{P}_c)\|_2^2$, where black pixels correspond to the minimal and white pixels to the maximal error.

$\mathbb{R}^{2m}$. Formally the desired partial derivatives are given as the Jacobian matrix

$$\mathbf{J}_{c,t} = \frac{\partial G_c}{\partial \mathfrak{X}_t} = \frac{\partial G_c}{\partial(\check{x}_{1,1}, \check{x}_{1,2}, \dots \check{x}_{m,2})} = \begin{pmatrix} \nabla^{\intercal} G_{c,j} \\ \vdots \\ \nabla^{\intercal} G_{c,R} \end{pmatrix} \tag{5.10}$$

with R being the number of depth pixels and the row vector $\nabla^{\intercal} G_{c,j}$ being the gradient of the depth value of the j-th pixel, computed by the rendering function. However, by virtue of discontinuities typically occurring at the boundary of objects due to occlusion, the rendering pipeline is in general not differentiable. To overcome this challenge we employ the approximative differentiable renderer framework OpenDR [68]. OpenDR approximates the partial derivatives of the rendering pipeline with respect to some input variables of the rendering process. By approximating the Jacobian matrices $\mathbf{J}_{c,t}$ with OpenDR and using automatic differentiation (cf. [44]) for the other terms, one can use of-the-shelf NLLSQ solver implementations to optimize the objective defined in (5.8). In Fig. 5.2 the gradient based optimization progress for a single multi-view frame in terms of residual images is depicted.

A drawback of the proposed gradient based approach is the dependence of the optimization result on the initialization. If the initial scene configuration corresponds to generated depth images where the individual renderings do not have any overlap with the observed image data, the gradients for those individuals will be zero. In consequence the optimization potentially gets stuck in a local minimum around the initialization state. Fig. 5.1(a) depicts an example where the initial 3D person renderings share some overlap with the observations. We soften the dependence on a particular initialization by applying a typical coarse to fine strategy. For an improved convergence behavior, the data error term is calculated for six layers of a Gaussian image pyramid.

While the coarse to fine strategy improves the convergence behavior, the proposed MAP approach still depends on a good initialization. Moreover, for inference prior knowledge about the number of people present in the scene is assumed (cf. Sect. 4.3). Both arguments emphasize that the MAP approach is suited for the refinement of a pre-existing scene configuration initialization, rather than providing people detection from scratch. In the evaluation (Sect. 6.6) it is shown that the proposed MAP method can be effectively used to fine-tune discrete detections obtained by mean-field variational inference (Sect. 5.2).

5.2 Mean-Field Variational Inference in Discrete Latent Space

In this section we introduce the main contribution of the present thesis. We propose approximate inference methods for the discrete scene configuration space introduced in Sect. 4.4. Based on the mean-field variational inference (MF-VI) method described in Sect. 3.1, we deduce the mean-field update equations to effectively approximate the desired probability distributions defined in Sect. 4.4. For better comprehensibility we first omit the temporal context and start with applying mean-field variational inference to the posterior for one time step (4.26) in Sect. 5.2.1. Building on these foundations, we extend the mean-field update equations in order to approximate the Bayesian filtering distribution (4.28). In Sect. 5.2.3, we finally deduce the mean-field update regulations in order to effectively approximate the full joint posterior distribution (4.27) of people present in the scene across space and time.

5.2.1 Data Posterior Distribution

In this section we propose a probabilistic inference method for the discrete posterior distribution $p(\mathbf{x}_t \mid \mathbf{o}_t)$ (4.26) at a single time step t. Inspired by POM [38] (cf. Sect. 3.2), we employ mean-field variational inference (cf. Sect. 3.1.3) to effectively approximate the desired discrete distribution $p(\mathbf{x}_t \mid \mathbf{o}_t)$ (4.26). First, we derive the corresponding mean-field update equations in detail. In Sect. 5.2.1.1 we introduce an approximation for the deduced update expectations, enabling an iterative mean-field optimization algorithm for real-world applications. Note that this section is an extension of work previously published [108, Ch. 3] by the author.

Since we only consider the observations at one time step in this section, we omit the time index t for ease of notation and restate the distribution (4.26) of a scene configuration $\mathbf{x}$ given the observations $\mathbf{o}$ as

$$p(\mathbf{x} \mid \mathbf{o}) = \frac{\prod_{c=1}^{C} p(o_c \mid \mathbf{x}) \prod_{i=1}^{n} p(x_i)}{\sum_{\mathbf{x}' \in \{0,1\}^n} \prod_{c=1}^{C} p\big(o_c \mid \mathbf{x}'\big) \prod_{i=1}^{n} p\big(x_i'\big)} \,. \tag{5.11}$$

Due to the high dimensional scene configuration space $\{0,1\}^n$, the partition function (evidence) in the denominator of (5.11) is intractable, and we cannot directly compute the posterior distribution. Instead, we apply Kullback-Leibler variational inference (cf. Sect. 3.1.2) to approximate the inconvenient distribution $p(\mathbf{x} \mid \mathbf{o})$ by a simpler proxy distribution $q(\mathbf{x})$. Following the reasoning given in Sect. 3.1.2, we propose to use the reverse KL-divergence. Thus the objective for optimizing $q(\mathbf{x})$ can be expressed as

$$\begin{aligned} \hat{q}(\mathbf{x}) &= \arg\min_q \mathrm{KL}(q(\mathbf{x}) \parallel p(\mathbf{x} \mid \mathbf{o})) \\ &= \arg\min_q \langle \log q(\mathbf{x}) - \log p(\mathbf{x} \mid \mathbf{o}) \rangle_{q(\mathbf{x})} \,. \end{aligned} \tag{5.12}$$

As elaborated in Sect. 3.1.3, a family of probability distributions (q-family) for the proxy distribution $q(\mathbf{x})$ has to be defined. According to Sect. 3.1.3 (3.19), we apply the naive mean-field assumption, which states that $q(\mathbf{x})$ is a product over its marginal probabilities

$$q(\mathbf{x}) = \prod_{i=1}^{n} q_i(x_i) \,. \tag{5.13}$$

Fig. 5.3(a) shows an exemplary mean-field state $q(\mathbf{x})$. Note that the mean-field assumption does not imply that the grid states $x_1, \ldots x_n$ in our model are assumed to be independent of each other. It is only a statement of the structure of the proxy distribution, which effects the iterative optimization schema. In order to deduce the mean-field update regulations a probability distribution over all $\mathbf{x}$ except of one single element x_i has to be defined (cf. Sect. 3.1). Let $q(\mathbf{x} \setminus x_i)$ denote the mean-field distribution excluding the element x_i,

$$q(\mathbf{x} \setminus x_i) = \prod_{j=1:j\neq i}^{n} q_j(x_j)\,. \tag{5.14}$$

According to Sect. 3.1.3 (3.26), the general mean-field equations

$$q_i(x_i) \propto \exp\Big(\langle \log p(\mathbf{x} \mid \mathbf{o})\rangle_{q(\mathbf{x}\setminus x_i)}\Big) \tag{5.15}$$

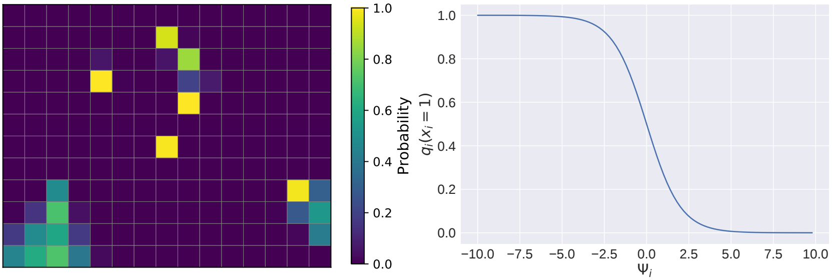
update a single marginal distribution $q_i(x_i)$ depending on the previous mean-field state $q(\mathbf{x} \setminus x_i)$. In Sect. 3.1.3 it is deduced that updating $q_i(x_i)$ asynchronously according to (5.15) will decrease the KL divergence in (5.12) (see also Barber [11, 625 ff.]). Since each x_i is Bernoulli distributed, (5.15) for x_i being in state 1, can be written as

$$q_i(x_i = 1) = \frac{1}{Z_i} \exp\Big(\langle \log p(\mathbf{o}, \mathbf{x} \mid x_i = 1)\rangle_{q(\mathbf{x}\setminus x_i)}\Big) \tag{5.16}$$

with the partition function

$$Z_i = \sum_{s\in\{0,1\}} \exp\Big(\langle \log p(\mathbf{o}, \mathbf{x} \mid x_i = s)\rangle_{q(\mathbf{x}\setminus x_i)}\Big)\,. \tag{5.17}$$

Considering that the posterior is proportional to the joint distribution $p(\mathbf{x} \mid \mathbf{o}) \propto p(\mathbf{o}, \mathbf{x})$, we could substitute the posterior distribution $p(\mathbf{x} \mid \mathbf{o})$ in the inner expectation of (5.15) by the joint distribution $p(\mathbf{o}, \mathbf{x})$, while the



(a) $q(\mathbf{x})$

(b)

Figure 5.3 (a) Exemplary mean-field distribution $q(\mathbf{x})$, where each grid cell corresponds to one marginal probability $q_i(x_i = 1)$. (b) Mean-field update for one marginal distribution $q_i(x_i = 1)$ (5.19) as a function of the expected value Ψ_i.

normalization by Z_i (5.17) ensures that (5.16) still satisfies the requirements of a probability mass function.

Applying the equality $\frac{e^x}{e^x+e^y} = \frac{1}{1+e^{y-x}}$ to (5.16) leads to the simplification

$$\begin{aligned} q_i(x_i = 1) = \Bigg[1 + \exp\Bigg(& \langle \log p(\mathbf{o}, \mathbf{x} \mid x_i = 0) \rangle_{q(\mathbf{x} \setminus x_i)} \\ & - \langle \log p(\mathbf{o}, \mathbf{x} \mid x_i = 1) \rangle_{q(\mathbf{x} \setminus x_i)} \Bigg)\Bigg]^{-1} . \end{aligned} \tag{5.18}$$

Using the linearity of expectation and the quotient rule of the logarithm we can rearrange (5.18) to

$$q_i(x_i = 1) = \left[1 + \exp\left(\underbrace{\left\langle \log \frac{p(\mathbf{o}, \mathbf{x} \mid x_i = 0)}{p(\mathbf{o}, \mathbf{x} \mid x_i = 1)} \right\rangle_{q(\mathbf{x} \setminus x_i)}}_{\Psi_i} \right)\right]^{-1} , \tag{5.19}$$

with Ψ_i being the mean-field expectation related to a marginal distribution $q_i(x_i = 1)$.

In Fig. 5.3(b) the mean-field update for one marginal distribution $q_i(x_i = 1)$ is plotted as a function of the expectation Ψ_i. Studying (5.19) and Fig. 5.3(b) already reveals the basic idea of an iterative mean-field update. The update of the probability of a person being present at location u_i (hence $x_i = 1$) increases as the expected value of the log ratio $\frac{p(\mathbf{o},\mathbf{x}|x_i=0)}{p(\mathbf{o},\mathbf{x}|x_i=1)}$ decreases. Considering the expectation with respect to the current mean-field state $q(\mathbf{x} \setminus x_i)$, the numerator reflects the probability of the grid location u_i being not occupied by a person, while the denominator reflects the probability of the grid location u_i being occupied. The correlation between the binary states of a scene configuration is taken into account as the expectation with respect to the current mean-field state $q(\mathbf{x} \setminus x_i)$, which evolves over optimization iterations. Informally, (5.19) answers the following question: Considering the observations $\mathbf{o}$ and the current mean-field state, is the presence of a person at location u_i more likely than the non-presence?

Given this basic intuition for an update of a distribution $q(\mathbf{x} \setminus x_i)$, we further continue with formally deducing the final mean-field equations. Inserting the probabilistic model defined in (4.16) and (4.26), the unconditioned expectation in (5.19) can be expressed as

$$\begin{aligned}\langle \log p(\mathbf{o},\mathbf{x})\rangle_{q(\mathbf{x}\setminus x_i)} &= \langle \log p(\mathbf{o} \mid \mathbf{x}) p(\mathbf{x})\rangle_{q(\mathbf{x}\setminus x_i)} \\ &= \left\langle -\sum_{c=1}^{C} \underbrace{\frac{1}{2\sigma_{\text{obs}}^2} \|o_c - G_c(\mathbf{x})\|_2^2}_{\delta(o_c, G_c(\mathbf{x}))} + \log p(\mathbf{x}) \right\rangle_{q(\mathbf{x}\setminus x_i)} \\ &= -\sum_{c=1}^{C} \langle \delta(o_c, G_c(\mathbf{x}))\rangle_{q(\mathbf{x}\setminus x_i)} + \langle \log p(\mathbf{x})\rangle_{q(\mathbf{x}\setminus x_i)}\,,\end{aligned} \tag{5.20}$$

with the image similarity function

$$\delta(I_1, I_2) = \frac{1}{2\sigma_{\text{obs}}^2} ||I_1 - I_2||_2^2\,. \tag{5.21}$$

Conditioning (5.20) on $x_i = \{0, 1\}$, inserting into (5.19) and making use of the linearity of expectation allows us to restate the final asynchronous mean-field update as

$$q_i(x_i = 1) = \left[1 + \exp\left(\tau_i + \sum_{c=1}^{C} \Psi_{c,i} \right) \right]^{-1} , \tag{5.22}$$

with the data expectation for one sensor given as

$$\Psi_{c,i} = \langle \delta(o_c, G_c(\mathbf{x}|x_i{=}1)) - \delta(o_c, G_c(\mathbf{x}|x_i{=}0)) \rangle_{q(\mathbf{x} \setminus x_i)} , \tag{5.23}$$

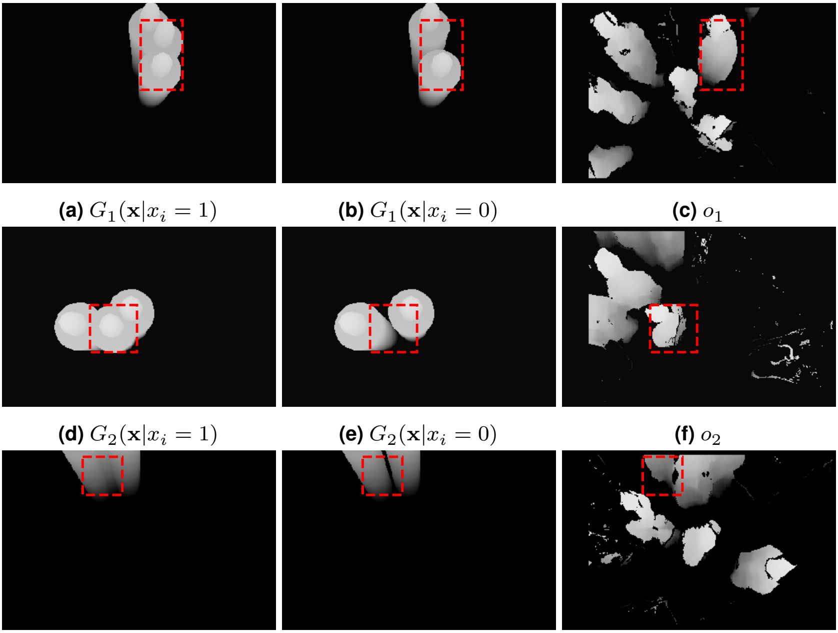
as well as the prior term τ_i. Based on the independence assumption $p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i)$ in (5.11) the prior term simplifies to

$$\tau_i = \left\langle \log \frac{p(\mathbf{x} \mid x_i = 0)}{p(\mathbf{x} \mid x_i = 1)} \right\rangle_{q(\mathbf{x} \setminus x_i)} = \log \frac{1 - p(x_i = 1)}{p(x_i = 1)} . \tag{5.24}$$

Note that $G_c(\mathbf{x}|x_i = 1)$ maps a scene configuration $\mathbf{x}$ to a synthetic depth image in the perspective of sensor $\mathcal{S}_c$ with x_i forced to 1 (cf. Fig. 5.4).

Fig. 5.4 depicts synthetic images and observations of three sensors, compared in expectation $\Psi_{c,i}$ (5.23). For the calculation of $\Psi_{c,i}$ these synthetic images need to be generated and compared with the observations for every possible scene configuration state. The images in Fig. 5.4 correspond to an exemplary scene configuration $\mathbf{x}$. Studying one row 5.4 (a-c) corresponding to sensor $\mathcal{S}_1$, the expectation (5.23) gets more accessible. For a given scene configuration $\mathbf{x}$, weighted with the current mean-field state $q(\mathbf{x} \setminus x_i)$, the observation o_1 is compared with a synthetic image with x_i forced to one ($G_1(\mathbf{x}|x_i = 1)$) and then compared with x_i forced to zero ($G_1(\mathbf{x}|x_i = 0)$). The difference of both image comparisons reflects the likeliness of a person occupying cell u_i under the scene configuration $\mathbf{x}$ and given the observation o_1.

Following the argument given in Fleuret *et al.* [38], one can also see how occlusion is handled implicitly: If the forward-model projection of



(a) $G_1(\mathbf{x}|x_i = 1)$ **(b)** $G_1(\mathbf{x}|x_i = 0)$ **(c)** o_1

(d) $G_2(\mathbf{x}|x_i = 1)$ **(e)** $G_2(\mathbf{x}|x_i = 0)$ **(f)** o_2

(g) $G_3(\mathbf{x}|x_i = 1)$ **(h)** $G_3(\mathbf{x}|x_i = 0)$ **(i)** o_3

Figure 5.4 Illustration of synthetic images and observations compared in the expectation $\Psi_{c,i}$ for one scene configuration $\mathbf{x}$. Each row corresponds to one sensor $\mathcal{S}_c$. The red dashed rectangle illustrates the bounding box corresponding to the rendering of a person present at location u_i. The scene configuration $\mathbf{x}$ shown here is zero for every grid location except for two neighbors of u_i.

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 $\delta(o_c, G_c(\mathbf{x}|x_i = s))$. Thus, the expectation $\Psi_{c,i}$ in (5.23) converges to zero and the corresponding marginal distribution q_i equals the prior.

5.2.1.1 Approximate Mean-Field Update

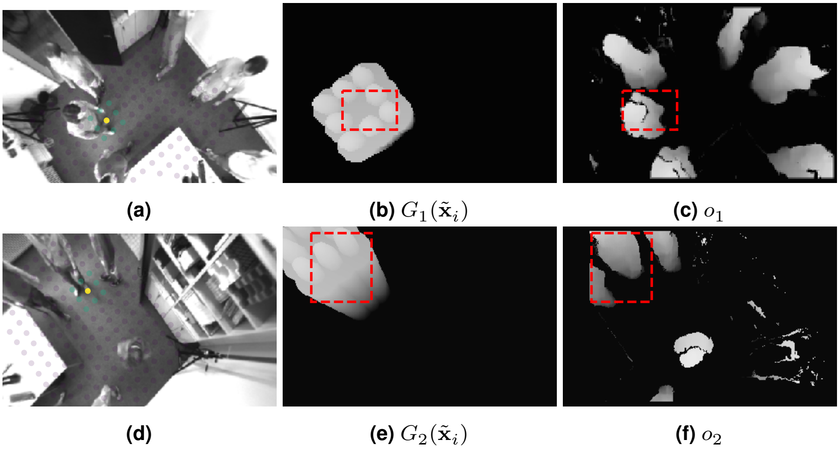
In spite of the mean-field assumption, (5.22) is still intractable due to the expectation $\Psi_{c,i} = \langle\cdot\rangle_{q(\mathbf{x}\setminus x_i)}$ (5.23), which requires to calculate a sum over all possible 2^{n-1} scene configuration states $\mathbf{x} \setminus x_i \in \{0,1\}^{n-1}$. We approximate the expected value $\Psi_{c,i}$ considering only the relevant subset of scene configurations. Therefore, we exploit the fact that the difference

$$\delta(o_c, G_c(\mathbf{x}|x_i = 1)) - \delta(o_c, G_c(\mathbf{x}|x_i = 0)) \tag{5.25}$$

only depends on the pixels belonging to the silhouette of the projection of the 3D model at location u_i (cf. Fig. 5.4). For a simpler and faster implementation, we approximate the region of belonging pixels by the corresponding axis-aligned rectangular bounding boxes, given as $I_c[u_i]$ (red rectangle in Fig. 5.4). Only scene configurations, for which the pixel values inside the bounding box $I_c[u_i]$ of the generated image $G_c(\mathbf{x})$ are affected, need to be evaluated for the expectation $\Psi_{c,i}$ in (5.22). We assume that only the projections of the direct eight neighbors of a grid location u_i intersect with the bounding box $I_c[u_i]$ (cf. Fig. 5.5). 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 $\Psi_{c,i}$ by the reduced neighborhood scene configuration $\tilde{\mathbf{x}}_i \in \{0,1\}^8$. Since the local neighborhood (including x_i) allows only $2^9 = 512$ possible local scene configurations, the expectation can be approximated efficiently.

Instead of the image distance $\delta(\cdot,\cdot)$ derived from the data likelihood in Sect. 4.2.2, 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 in the mean-field optimization, 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

$$\begin{aligned}\delta_{\text{asym}}(o,g) = {} & \alpha \left\| o \odot \overline{M}(g) \right\|_1 + (2-\alpha) \left\| g \odot \overline{M}(o) \right\|_1 \\ & + \left\| (o-g) \odot M(o) \odot M(g) \right\|_1 ,\end{aligned} \tag{5.26}$$


(a) **(b)** $G_1(\tilde{\mathbf{x}}_i)$ **(c)** o_1

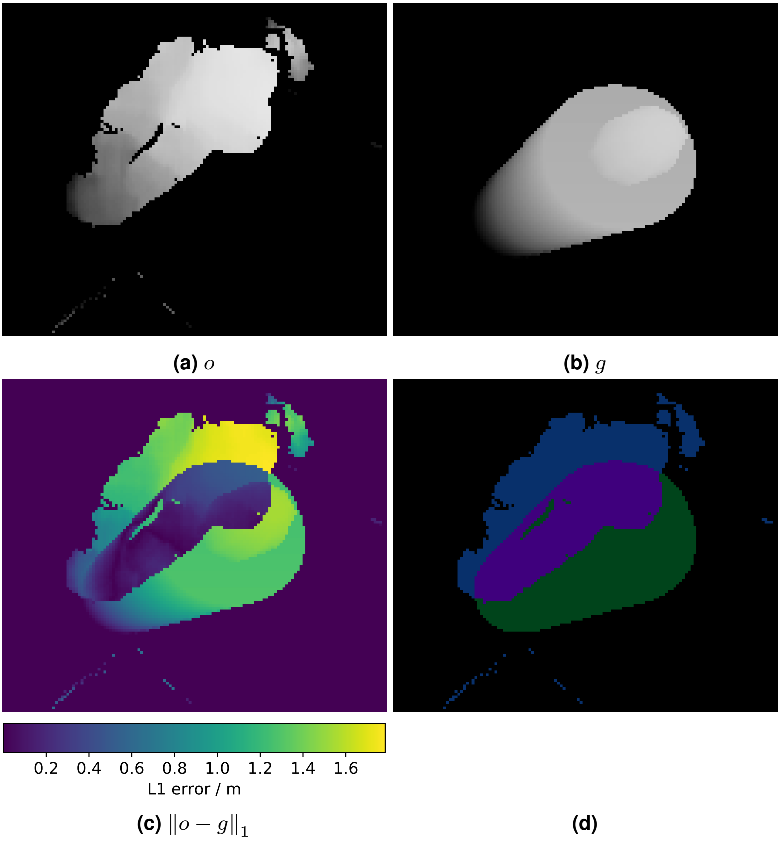
(d) **(e)** $G_2(\tilde{\mathbf{x}}_i)$ **(f)** o_2

Figure 5.5 Illustration of direct neighborhood of a grid cell u_i for sensor 1 (a-c) and sensor 2 (d-f). (a,d) show the direct grid neighbors in turquoise and the grid location u_i in yellow, respectively. (b,e) show the synthetic images where every element in the neighborhood scene configuration $\tilde{\mathbf{x}}_i$ is set to one. (c,f) show the corresponding depth observations. The red dashed rectangle shows the bounding box $I_c[u_i]$ corresponding to the rendering of a person present at location u_i in sensor $\mathcal{S}_c$.

with the design parameter $\alpha \in [0, 2]$.

In Fig. 5.6 the asymmetric image similarity is graphically illustrated, including the masks of three disjoint cases $o \odot \overline{M}(g)$, $g \odot \overline{M}(o)$ and $M(o) \odot M(g)$, respectively. 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 $o \odot \overline{M}(g)$ will be penalized more strongly (cf. blue mask in Fig. 5.6(d)). Let further

$$\delta_{x_i=s} = \frac{1}{2\sigma^2_{\text{obs}}} \delta_{\text{asym}}(o_c[u_i], G_c(\tilde{\mathbf{x}}_i | x_i = s)[u_i]) \tag{5.27}$$



(a) o

(b) g

(c) $\|o - g\|_1$

(d)

Figure 5.6 Illustration of different error terms occurring in the asymmetric image similarity. (a,b) show an exemplary observation and synthetic depth image used as input for the image similarity $\delta_{\text{asym}}(o, g)$ (5.26). (c) depicts the L1-norm residual image, where purple corresponds to a low error and yellow to a high error, respectively. (d) shows the masks used in (5.26) as colored overlay, whereas blue corresponds to $o \odot \overline{M}(g)$, green to $g \odot \overline{M}(o)$ and purple to $M(o) \odot M(g)$, respectively.

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

$$\tilde{\Psi}_{c,i} = \frac{1}{|I_c[u_i]|} \left\langle \delta_{x_i=1} - \delta_{x_i=0} \right\rangle_{q(\tilde{\mathbf{x}}_i)} . \tag{5.28}$$

Additionally, we normalize the expectation with respect to the size (number of pixels) of the image slice $|I_c[u_i]|$, to account for the viewpoint dependent size of a bounding box. In order to efficiently compute (5.28), we propose to pre-build for each u_i a visual dictionary[2] of image slices $I_c[u_i]$ for all 512 possible local scene configurations $\tilde{\mathbf{x}_i}$.

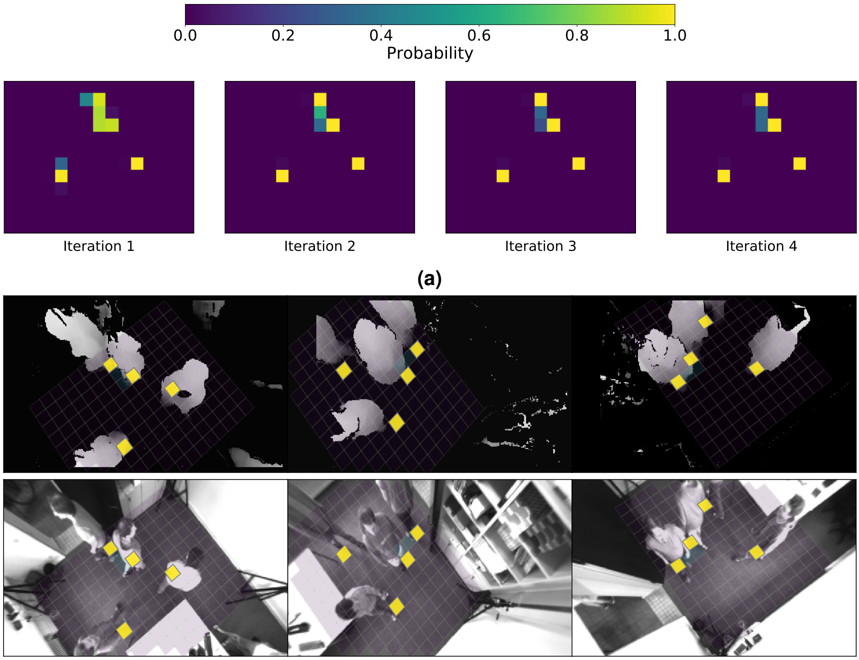
5.2.1.2 Optimization Details

According to Sect. 3.1.4 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 $q(\mathbf{x} \setminus x_i)$. To avoid oscillating effects during the optimization we use the asynchronous coordinate-ascent variational inference (CAVI) method as listed in Algorithm 1 in Sect. 3.1.4. Hence, the probability for each $q_i(x_i)$ will be updated asynchronously with respect to the previous mean-field state $q(\mathbf{x} \setminus x_i)$ according to the final update equation

$$q_i(x_i = 1) = \left[1 + \exp\left(\tau_i + \sum_{c=1}^{C} \tilde{\Psi}_{c,i} \right) \right]^{-1} . \tag{5.29}$$

An exemplary CAVI mean-field optimization with four iterations is depicted in Fig. 5.7.

[2] In this context the visual dictionary refers to a pre-computed list of synthetic depth image slices, enabling efficient computation of the expectation $\tilde{\Psi}_{c,i}$.



(a)

(b)

Figure 5.7 Mean-field results for one exemplary multi-view frame. (a) shows mean-field state $q(\mathbf{x})$, given as a probability map of the marginals, at every iteration. (b) illustrates the final mean-field optimization results $\hat{q}(\mathbf{x})$, back projected into each sensor view as probability maps.

5.2.2 Bayesian Filtering

In this section we extend the mean-field approach introduced in the previous section by taking the history of consecutive observations into account. As introduced in the previous section we use the mean-field assumption to approximate the filtering distribution. However, instead of just approximating the data posterior $p(\mathbf{x}_t \mid \mathbf{o}_t)$ at a single time step, in this section we approximate the Bayesian filtering distribution $p(\mathbf{x}_t \mid \mathbf{o}_{1:t})$ (4.28), incorporating the previous observations. Since the desired distribution is defined recursively (cf. (4.11)), the past state is condensed in the previous filtering distribution $p(\mathbf{x}_{t-1} \mid \mathbf{o}_{1:t-1})$. In practice, we use the final result from the last mean-field optimization $\hat{q}(\mathbf{x}_{t-1})$ as an approximation of the previous filtering distribution. In consequence, each mean-field update directly depends on the approximated marginal distributions $\hat{q}(\mathbf{x}_{t-1})$ from the previous time step. Following Sect. 5.2.1 the objective is to minimize the KL-divergence between the desired distribution $p(\mathbf{x}_t \mid \mathbf{o}_{1:t})$ and a proxy distribution $q(\mathbf{x}_t)$,

$$\hat{q}(\mathbf{x}_t) = \arg\min_q \mathrm{KL}(q(\mathbf{x}_t) \parallel p(\mathbf{x}_t \mid \mathbf{o}_{1:t})) \,. \tag{5.30}$$

Following Sect. 5.2.1 we apply the naive mean-field assumption (5.13), assuming the factorization $q(\mathbf{x}_t) = \prod_{i=1}^{n} q_i(x_i)$.

Using the general mean-field equation and applying the transformations presented in (5.15)–(5.19), the mean-field equations for the filtering distribution are given as

$$q_i(x_i = 1) = \left[1 + \exp\left(\underbrace{\left\langle \log \frac{p(\mathbf{o}_{1:t}, \mathbf{x}_t \mid x_i = 0)}{p(\mathbf{o}_{1:t}, \mathbf{x}_t \mid x_i = 1)} \right\rangle_{q(\mathbf{x}_t \backslash x_i)}}_{\Psi_i^{\text{filter}}} \right) \right]^{-1} . \tag{5.31}$$

Inserting the probabilistic model from (4.28), using the linearity of expectation and the equality $\log(\frac{ab}{cd}) = \log\frac{a}{c} + \log\frac{b}{d} \,:\, a, b, c, d > 0$ allows the separation of the filtering expectation Ψ_i^{filter} into two disjoint parts; the already known data expectation and a predictive expectation

$$\Psi_i^{\text{filter}} = \Psi_i^{\text{data}} + \underbrace{\left\langle \log \frac{p(\mathbf{x}_t \mid \mathbf{o}_{1:t-1}, x_i = 0)}{p(\mathbf{x}_t \mid \mathbf{o}_{1:t-1}, x_i = 1)} \right\rangle_{q(\mathbf{x}_t \backslash x_i)}}_{\Psi_i^{\text{pred}}} . \tag{5.32}$$

The data term for the current time step is the same as defined in (5.22), thus given as

$$\Psi_i^{\text{data}} = \sum_{c=1}^{C} \langle \delta(o_c, G_c(\mathbf{x}|x_i{=}1)) - \delta(o_c, G_c(\mathbf{x}|x_i{=}0)) \rangle_{q(\mathbf{x}\setminus x_i)} \,. \tag{5.33}$$

The dependency on observations from previous time steps $\mathbf{o}_{1:t-1}$ is condensed in the recursively defined predictive distribution $p(\mathbf{x}_t \mid \mathbf{o}_{1:t-1})$. After inserting the predictive distribution (4.29) and applying the proposed factorization of the dynamics model (4.30), the expectation Ψ_i^{pred} expands to

$$\begin{aligned} \Psi_i^{\text{pred}} &= \left\langle \log \frac{\sum_{\mathbf{x}_{t-1}} p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, x_i = 0) p(\mathbf{x}_{t-1} \mid \mathbf{o}_{1:t-1})}{\sum_{\mathbf{x}_{t-1}} p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, x_i = 1) p(\mathbf{x}_{t-1} \mid \mathbf{o}_{1:t-1})} \right\rangle_{q(\mathbf{x}_t \setminus x_i)} \\ &= \left\langle \log \frac{\sum_{\mathbf{x}_{t-1}} \prod_{j \in N_i'} p\big(x_{j,t} \mid \mathbf{x}_{t-1}, x_i = 0\big) p(\mathbf{x}_{t-1} \mid \mathbf{o}_{1:t-1})}{\sum_{\mathbf{x}_{t-1}} \prod_{j \in N_i'} p\big(x_{j,t} \mid \mathbf{x}_{t-1}, x_i = 1\big) p(\mathbf{x}_{t-1} \mid \mathbf{o}_{1:t-1})} \right\rangle_{q(\mathbf{x}_t \setminus x_i)} , \end{aligned} \tag{5.34}$$

with $\sum_{\mathbf{x}_{t-1}}$ being the shorthand for $\sum_{\mathbf{x}_{t-1} \in \{0,1\}^n}$ and $N_i' = N_i \cup i$ the index set of the direct neighborhood of the cell u_i, including the index i. As a direct consequence of the chosen transition function $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ (4.40), which assumes that a person moves only in the direct neighborhood in one time step, the set of scene configurations can be reduced to the local neighborhood scene configurations $\tilde{\mathbf{x}_i}$ of the cell u_i. Therefore, the sum in (5.34) can be replaced by $\sum_{\tilde{\mathbf{x}}_{i,t-1} \in \{0,1\}^9}$, which makes the computation of the predictive expectation feasible. The filtering distribution of the last time $p(\mathbf{x}_{t-1} \mid \mathbf{o}_{1:t-1})$ states the likeliness of a previous state $\mathbf{x}_{t-1}$. In practice this distribution is approximated by the result of the last mean-field optimization $\hat{q}(\mathbf{x}_{t-1})$. Under this assumption and after applying the definition of expectation to the nominator and denominator of (5.34), the approximated predictive expectation can be written as

$$\tilde{\Psi}_i^{\text{pred}} = \left\langle \log \frac{\left\langle \prod_{j \in N_i'} p\big(x_{j,t} \mid \mathbf{x}_{t-1}, x_i = 0\big) \right\rangle_{\hat{q}(\mathbf{x}_{t-1})}}{\left\langle \prod_{j \in N_i'} p\big(x_{j,t} \mid \mathbf{x}_{t-1}, x_i = 1\big) \right\rangle_{\hat{q}(\mathbf{x}_{t-1})}} \right\rangle_{q(\mathbf{x}_t \setminus x_i)} . \tag{5.35}$$

Informally, the influence of (5.34) on a mean-field update (5.31) can be interpreted as follows: considering the distribution of the previous state $\mathbf{x}_{t-1}$, the proposed dynamics model $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$, and the current mean-field state $q(\mathbf{x}_t \setminus x_i)$, is the presence of a person at location u_i more likely than the non-presence?

The final mean-field optimization for the Bayesian filtering distribution is implemented analogous to the data distribution elaborated in Sect. 5.2.1.2.

5.2.3 Temporal Smoothing

Building upon the ideas introduced in the previous sections we address the approximation of the full posterior distribution $p(\mathbf{x}_{1:T} \mid \mathbf{o}_{1:T})$ in this section. In contrast to Sect. 5.2.1 and Sect. 5.2.2, where the marginal probabilities for one single time step t are approximated, we now aim for a joint approximation of the distribution of people present in the scene for a sequence of time steps. As a consequence, we extend the naive mean-field assumption in (5.13) to a factorization over a sequence of scene configurations $\mathbf{x}_{1:T}$ and deduce the corresponding mean-field update equations. Notice that the following section is an extension of previously published work [109, Ch. 3.B] by the author.

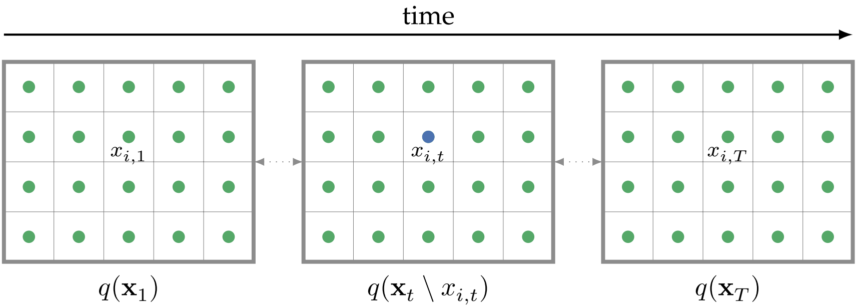


Figure 5.8 Illustration of the fully factorized temporal mean-field distribution $q(\mathbf{x}_{1:T} \setminus x_{i,t})$. The temporal mean-field distribution can be separated into T spatial mean-fields $q(\mathbf{x}_k)$. For an update of a single marginal distribution $q(x_{i,t})$ the expectation is taken with respect to all other spatio-temporal nodes.

Following the previous sections, we use mean-field variational inference to approximate the distribution $p(\mathbf{x}_{1:T} \mid \mathbf{o}_{1:T})$ by a simpler proxy distribution $q(\mathbf{x}_{1:T})$. The optimization objective is given as

$$\hat{q}(\mathbf{x}_{1:T}) = \arg\min_q \mathrm{KL}(q(\mathbf{x}_{1:T}) \parallel p(\mathbf{x}_{1:T} \mid \mathbf{o}_{1:T})) \tag{5.36}$$

$$= \arg\min_q \langle \log q(\mathbf{x}_{1:T}) - \log p(\mathbf{x}_{1:T} \mid \mathbf{o}_{1:T}) \rangle_{q(\mathbf{x}_{1:T})} . \tag{5.37}$$

In order to perform joint inference for a sequence of scene configurations $\mathbf{x}_{1:T}$ a structure for the proxy distribution $q(\mathbf{x}_{1:T})$ enabling a computationally feasible approximation needs to be defined. We extend the mean-field assumption (5.13) by additionally factorizing over time

$$q(\mathbf{x}_{1:T}) = \prod_{i=1}^{n} \prod_{t=1}^{T} q_{i,t}(x_{i,t}) , \tag{5.38}$$

where each $q_{i,t}(x_{i,t})$ denotes the marginal probability distribution of a person present at location u_i at time step t (cf. Fig. 5.8). Analogous to (5.14) let

$$q(\mathbf{x}_{1:T} \setminus x_{i,t}) = \prod_{j=1}^{n} \prod_{\substack{k=1 \\ k \neq t \vee j \neq i}}^{T} q_{j,k}(x_{j,k}) \tag{5.39}$$

be the mean-field distribution without the element $x_{i,t}$ (cf. Fig. 5.8).

According to the general mean-field equation (cf. Sect. 3.1) the optimal update with respect to the objective (5.37) is given as

$$q_{i,t}(x_{i,t}) \propto \exp\Big(\langle \log p(\mathbf{x}_{1:T} \mid \mathbf{o}_{1:T}) \rangle_{q\left(\mathbf{x}_{1:T} \setminus x_{i,t}\right)} \Big) \,. \tag{5.40}$$

Re-arranging as in (5.15)–(5.19), the final update for $x_{i,t}$ being in state one is given as

$$q_{i,t}(x_{i,t} = 1) = \left[1 + \exp\left(\Psi_{i,t}^{\text{smooth}} \right) \right]^{-1} \,. \tag{5.41}$$

Inserting the probabilistic model defined in (4.27) and using the relation $\log(\frac{ab}{cd}) = \log \frac{a}{c} + \log \frac{b}{d} \,:\, a, b, c, d > 0$, the expectation $\Psi_{i,t}^{\text{smooth}}$ in (5.41) expands to

$$\begin{aligned} \Psi_{i,t}^{\text{smooth}} &= \left\langle \log \frac{p\big(\mathbf{o}_{1:T}, \mathbf{x}_{1:T} \bigm| x_{i,t} = 0\big)}{p\big(\mathbf{o}_{1:T}, \mathbf{x}_{1:T} \bigm| x_{i,t} = 1\big)} \right\rangle_{q\left(\mathbf{x}_{1:T} \setminus x_{i,t}\right)} && (5.42) \\ &= \left\langle \log \frac{\prod_{k=1}^{T} p\big(\mathbf{o}_k \bigm| \mathbf{x}_k, x_{i,t} = 0\big)}{\prod_{k=1}^{T} p\big(\mathbf{o}_k \bigm| \mathbf{x}_k, x_{i,t} = 1\big)} \right. && (5.43) \\ &\qquad \left. + \log \frac{\prod_{k=1}^{T} p\big(\mathbf{x}_k \bigm| \mathbf{x}_{k-1}, x_{i,t} = 0\big)}{\prod_{k=1}^{T} p\big(\mathbf{x}_k \bigm| \mathbf{x}_{k-1}, x_{i,t} = 1\big)} \right\rangle_{q\left(\mathbf{x}_{1:T} \setminus x_{i,t}\right)} \,. \end{aligned}$$

Using the linearity of expectation, (5.43) can be expressed as the sum of a data and a temporal expectation

$$\Psi_{i,t}^{\text{smooth}} = \Psi_{i,t}^{\text{data}} + \Psi_{i,t}^{\text{temp}} \,. \tag{5.44}$$

Studying the expression in (5.43) reveals that all terms in (5.43) independent of $x_{i,t}$ cancel out, since the nominator and denominator for the factors independent of $x_{i,t}$ are identical. Hence, the data term can be isolated to

$$\Psi_{i,t}^{\text{data}} = \left\langle \log \frac{\prod_{k=1}^{T} p\big(\mathbf{o}_k \mid \mathbf{x}_k, x_{i,t} = 0\big)}{\prod_{k=1}^{T} p\big(\mathbf{o}_k \mid \mathbf{x}_k, x_{i,t} = 1\big)} \right\rangle_{q\left(\mathbf{x}_{1:T} \backslash x_{i,t}\right)} \tag{5.45}$$

$$= \left\langle \log \frac{p\big(\mathbf{o}_t \mid \mathbf{x}_t, x_{i,t} = 0\big)}{p\big(\mathbf{o}_t \mid \mathbf{x}_t, x_{i,t} = 1\big)} \right\rangle_{q(\mathbf{x}_t \backslash x_i)} . \tag{5.46}$$

In consequence the data term does not depend on the temporal context anymore and is identical to the data term derived in (5.22). According to (5.43) and (5.44), the temporal expectation is given as

$$\Psi_{i,t}^{\text{temp}} = \left\langle \log \frac{\prod_{k=1}^{T} p\big(\mathbf{x}_k \mid \mathbf{x}_{k-1}, x_{i,t} = 0\big)}{\prod_{k=1}^{T} p\big(\mathbf{x}_k \mid \mathbf{x}_{k-1}, x_{i,t} = 1\big)} \right\rangle_{q\left(\mathbf{x}_{1:T} \backslash x_{i,t}\right)} . \tag{5.47}$$

Following the same argument as for the data term, all factors which are independent of the state $x_{i,t}$ cancel out. Depending on k, the forced state $x_{i,t}$ can be either part of the condition or the argument of the distribution $p\big(\mathbf{x}_k \mid \mathbf{x}_{k-1}, x_{i,t} = s\big)$. Analyzing the temporal expectation $\Psi_{i,t}^{\text{temp}}$ reveals that the factors in the product $\prod_{k=1}^{T} p\big(\mathbf{x}_k \mid \mathbf{x}_{k-1}, x_{i,t} = s\big)$ with $s \in \{0, 1\}$ can be separated into three disjoint cases, depending on the value of k:

1) If $k = t$, the corresponding factor is given as $p\big(\mathbf{x}_t \mid \mathbf{x}_{t-1}, x_{i,t} = s\big)$. Since $x_{i,t}$ is an element of the scene configuration $\mathbf{x}_t$, this distributions depend on the state of $x_{i,t}$.

2) If $k = t + 1$, the corresponding factor is given as $p\big(\mathbf{x}_{t+1} \mid \mathbf{x}_t, x_{i,t}\big)$. Here the condition $\mathbf{x}_t$ depends directly on the state of $x_{i,t}$.

3) For all other values of k with $k \neq t \wedge k \neq t + 1$, the distribution $p\big(\mathbf{x}_k \mid \mathbf{x}_{k-1}, \cancel{x_{i,t}}\big)$ is independent of $x_{i,t}$. In consequence those factors are identical for $x_{i,t} = 0$ and $x_{i,t} = 1$ and cancel out in (5.47).

Considering these findings and applying the linearity of expectation, we can separate the temporal expectation $\Psi_{i,t}^{\text{temp}}$ (5.47) into two parts, referred to as past and future expectation,

$$\begin{aligned}
\Psi_{i,t}^{\text{temp}} &= \left\langle \log \frac{p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, x_{i,t}=0)p(\mathbf{x}_{t+1} \mid \mathbf{x}_t, x_{i,t}=0)}{p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, x_{i,t}=1)p(\mathbf{x}_{t+1} \mid \mathbf{x}_t, x_{i,t}=1)} \right\rangle_{q(\mathbf{x}_{1:T} \setminus x_{i,t})} \\
&= \underbrace{\left\langle \log \frac{p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, x_{i,t}=0)}{p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, x_{i,t}=1)} \right\rangle_{q(\mathbf{x}_t \setminus x_{i,t}, \mathbf{x}_{t-1})}}_{\Psi_{i,t}^{\text{past}}} \qquad (5.48) \\
&\quad + \underbrace{\left\langle \log \frac{p(\mathbf{x}_{t+1} \mid \mathbf{x}_t, x_{i,t}=0)}{p(\mathbf{x}_{t+1} \mid \mathbf{x}_t, x_{i,t}=1)} \right\rangle_{q(\mathbf{x}_t \setminus x_{i,t}, \mathbf{x}_{t+1})}}_{\Psi_{i,t}^{\text{future}}}, \qquad (5.49)
\end{aligned}$$

where $\Psi_{i,t}^{\text{past}}$ corresponds to the past expectation and $\Psi_{i,t}^{\text{future}}$ corresponds to the future expectation. Based on this intermediate result, both expectations can be further simplified. Inserting the dynamics model (4.30) in (5.48) and considering that all terms which are independent of $x_{i,t}$ cancel out, the past expectation is given as

$$\begin{aligned}
\Psi_{i,t}^{\text{past}} &= \left\langle \log \frac{p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, x_{i,t}=0)}{p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, x_{i,t}=1)} \right\rangle_{q(\mathbf{x}_t \setminus x_{i,t}, \mathbf{x}_{t-1})} \qquad (5.50) \\
&= \left\langle \log \frac{\prod_{j \in N_i'} p(x_{j,t} \mid \mathbf{x}_{t-1}, x_{i,t}=0)}{\prod_{j \in N_i'} p(x_{j,t} \mid \mathbf{x}_{t-1}, x_{i,t}=1)} \right\rangle_{q(\mathbf{x}_t \setminus x_{i,t}, \mathbf{x}_{t-1})} \qquad (5.51) \\
&= \left\langle \log \frac{1 - p(x_{i,t}=1 \mid \mathbf{x}_{t-1})}{p(x_{i,t}=1 \mid \mathbf{x}_{t-1})} \right\rangle_{q(\mathbf{x}_{t-1})}. \qquad (5.52)
\end{aligned}$$

Finally, the past expectation only depends on the previous state $\mathbf{x}_{t-1}$, thus the expectation is only taken with respect to the distribution $q(\mathbf{x}_{t-1})$. Following the derivation of the past expectation, inserting the dynamics model (4.30) in (5.49) and canceling all terms independent of $x_{i,t}$, the future expectation simplifies to

$$\Psi_{i,t}^{\text{future}} = \left\langle \log \frac{p\big(\mathbf{x}_{t+1} \mid \mathbf{x}_t, x_{i,t} = 0\big)}{p\big(\mathbf{x}_{t+1} \mid \mathbf{x}_t, x_{i,t} = 1\big)} \right\rangle_{q(\mathbf{x}_t \backslash x_{i,t}, \mathbf{x}_{t+1})} \tag{5.53}$$

$$= \left\langle \log \frac{\prod_{j \in N_i'} p\big(x_{j,t+1} \mid \mathbf{x}_t, x_{i,t} = 0\big)}{\prod_{j \in N_i'} p\big(x_{j,t+1} \mid \mathbf{x}_t, x_{i,t} = 1\big)} \right\rangle_{q(\mathbf{x}_t \backslash x_{i,t}, \mathbf{x}_{t+1})} \tag{5.54}$$

$$= \left\langle \sum_{j \in N_i'} \log \frac{p\big(x_{j,t+1} \mid \mathbf{x}_t, x_{i,t} = 0\big)}{p\big(x_{j,t+1} \mid \mathbf{x}_t, x_{i,t} = 1\big)} \right\rangle_{q(\mathbf{x}_t \backslash x_{i,t}, \mathbf{x}_{t+1})} , \tag{5.55}$$

and depends on the current state $\mathbf{x}_t$ and the state $\mathbf{x}_{t+1}$ at the next time step.

Since the dynamics model (4.40) introduced in Sect. 4.4.1 operates only on a local neighborhood, we only need to consider the reduced neighborhood scene configurations $\tilde{\mathbf{x}}_{i,t} \in \{0,1\}^9$, making the estimation of the expectations (5.50) and (5.53) computationally feasible. Under the reduced scene state distributions, the final expectations used in the temporal mean-field optimization are given as the past expectation

$$\Psi_{i,t}^{\text{past}} = \left\langle \log \frac{1 - p\big(x_{i,t} = 1 \mid \tilde{\mathbf{x}}_{i,t-1}\big)}{p\big(x_{i,t} = 1 \mid \tilde{\mathbf{x}}_{i,t-1}\big)} \right\rangle_{q(\tilde{\mathbf{x}}_{i,t-1})} , \tag{5.56}$$

and future expectation

$$\Psi_{i,t}^{\text{future}} = \left\langle \sum_{j \in N_i'} \log \frac{p\big(x_{j,t+1} \mid \tilde{\mathbf{x}}_{i,t}, x_{i,t} = 0\big)}{p\big(x_{j,t+1} \mid \tilde{\mathbf{x}}_{i,t}, x_{i,t} = 1\big)} \right\rangle_{q(\tilde{\mathbf{x}}_{i,t}, \tilde{\mathbf{x}}_{i,t+1})} . \tag{5.57}$$

Inserting into (5.41), the final mean-field equations for the temporal smoothing approach are given as

$$q_{i,t}(x_{i,t}=1) = \left[1+\exp\left(\Psi^{\text{data}}_{i,t} + \underbrace{\Psi^{\text{past}}_{i,t} + \Psi^{\text{future}}_{i,t}}_{\Psi^{\text{temp}}_{i,t}}\right)\right]^{-1} . \tag{5.58}$$

5.2.3.1 Optimization Details

As in the previous sections, for the mean-field optimization we use CAVI (cf. Sect. 3.1.4). Thus all $q_{i,t}(\cdot)$ are updated sequentially according to (5.58) with respect to the previous mean-field state $q(\mathbf{x}_{1:T} \setminus x_{i,t})$. In Algorithm 2 the joint temporal mean-field optimization algorithm is listed. In each iteration the time slices $q(\mathbf{x}_t)$ are consecutively updated from $1, \ldots, T$. 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 (5.53) relies on the mean-field state from the next time step we disable the future term in the first iteration. Considering that people can enter the observable area, we initialize all border grid cells with a probability of 0.5. To weight the temporal terms we extend (5.44) to $\Psi^{\text{smooth}}_{i,t} = \Psi^{\text{data}}_{i} + \lambda_{\text{past}}\Psi^{\text{past}}_{i,t} + \lambda_{\text{future}}\Psi^{\text{future}}_{i,t}$ with $\lambda_{\text{past}}, \lambda_{\text{future}} \in [0,1]$.

5.3 CNN Inference

In the previous sections, inference is obtained by iterative optimization methods, explicitly derived from the probabilistic model introduced in Sect. 4. In this section a different method is applied for inference in discrete latent space. We propose an end-to-end multi-view convolutional neural network (CNN) architecture to approximate the marginal probabilities of people present in the scene for a single time step. In contrast to the explicit iterative mean-field optimization, the CNN framework is demanding at

Algorithm 2 Joint Temporal Mean-Field Optimization

```
procedure OPTIMIZETEMPORALMF
    $\hat{q}(\mathbf{x}_{1:T}) \leftarrow$ init()                                  ▷ init mean-field
    for all $m \in \{0, \dots, \texttt{ITERATIONS}\}$ do
        for all $t \in \{1, \dots, T\}$ do                        ▷ iterate along time axis
            for all $i \in \{1, \dots, n\}$ do                   ▷ iterate over all grid cells
                $\hat{q}_{i,t} \leftarrow \left[1 + \exp\left(\langle\cdot\rangle_{q(\tilde{\mathbf{x}}_{i,t-1}, \tilde{\mathbf{x}}_{i,t}, \tilde{\mathbf{x}}_{i,t+1})}\right)\right]^{-1}$

                $\hat{q}(\mathbf{x}_{1:T}) \leftarrow$ update($\hat{q}_{i,t}$)        ▷ asynchronous MF update
            end for
        end for
    end for
end procedure
```

training time, but once the network is trained, inference can be obtained by a single deterministic forward pass. For a fair comparison with the proposed mean-field methods (Sect. 5.2) and to overcome the lack of a domain-specific large scale data-set, we sample from the generative scene model introduced in Sect. 4.2.1 to generate synthetic training data. To narrow the gap between synthetic training images and real-world depth sensor observations, we extend the generative scene model to generate randomized synthetic training images (Sect. 5.3.2). In contrast to classical data-driven approaches the proposed multi-view CNN is only trained with synthetic depth images and does not rely on any real-world training data. Analogous to the mean-field inference for one time step, the network takes three foreground segmented depth images as input and predicts the marginal probability distributions of people present in the scene (cf. Fig. 5.9). Since the generative mean-field inference and the discriminative multi-view CNN framework only rely on the generative scene model, without further supervision, the methods can be directly compared at inference time. This section is an extension of previously published work [112, Ch.2] by the author.

Table 5.1 Parameters of CNN feature extraction.

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

5.3.1 End-to-End Multi-View CNN Architecture

In this section we introduce a multi-view CNN architecture for people detection in multiple depth images. For a fair comparison with the probabilistic methods, the same discrete grid model as proposed in Sect. 4.4 is used. Since we only consider the observations at one time step, we omit the time index t for ease of notation. Hence, a scene configuration is given as $\mathbf{x} = (x_1, \ldots, x_n)^\mathsf{T} \in \{0,1\}^n$ and the vector of foreground-segmented depth observations at one time step is stated as $\mathbf{o} = (o_1, \ldots, o_c)^\mathsf{T}$. The objective of the proposed end-to-end approach is to approximate the distribution

$$p(\mathbf{x} \mid \mathbf{o}) = \prod_{i=1}^{n} p(x_i \mid \mathbf{o}) \tag{5.59}$$

with $p(x_i \mid \mathbf{o})$ being the marginal probability of a person present at ground plane location u_i given the observations $\mathbf{o}$. To approximate (5.59) we propose a multi-view CNN architecture (cf. Fig. 5.9) which jointly exploits the depth observations from three sensors.

We observe no significant drop in performance when the input depth images are scaled-down by a factor of 2. Therefore we use input depth images of size 188×120 for each individual CNN-head, yielding the advantage that the GPU memory footprint can be reduced significantly. Generalization over the visual features is achieved by weights sharing across the input CNN-heads. The resulting feature maps of each CNN-

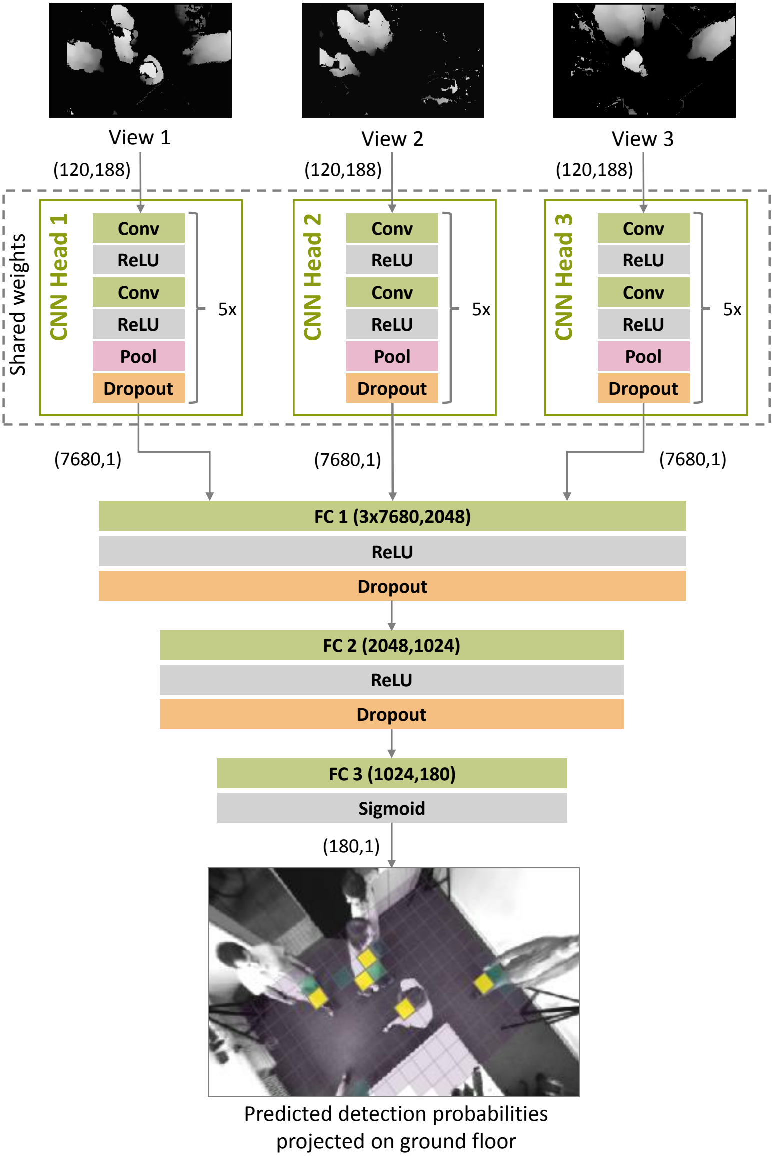


Figure 5.9 Overview of our proposed CNN multi-view architecture. Each input depth image serves as input for a CNN module. The output of the last fully connected layer predicts the marginal probability distribution of people present in the scene.

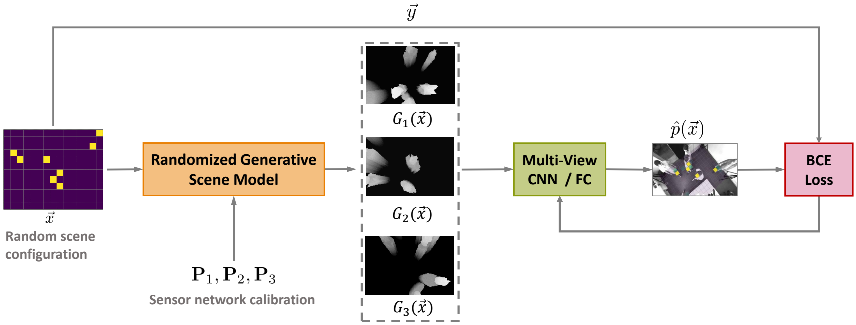

$\vec{y}$
$\vec{x}$
Random scene configuration
Randomized Generative Scene Model
$\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3$
Sensor network calibration
$G_1(\vec{x})$
$G_2(\vec{x})$
$G_3(\vec{x})$
Multi-View CNN / FC
$\hat{p}(\vec{x})$
BCE Loss

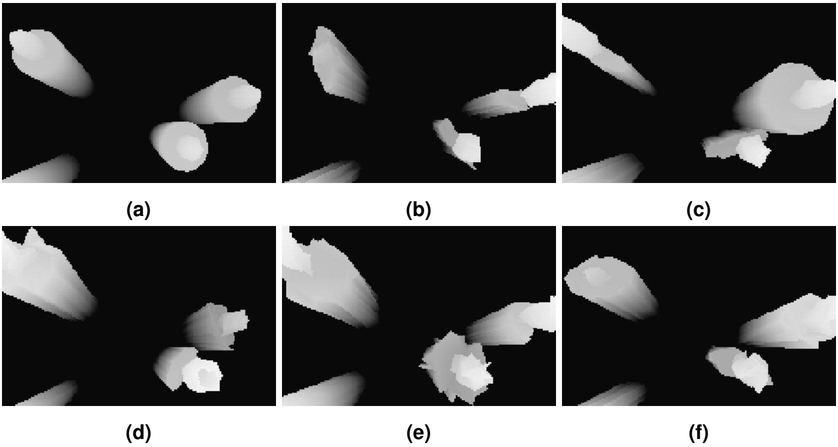
Figure 5.10 End-to-end multi-view training with randomized synthetic depth images.

head are concatenated and fed into a multilayer perceptron (MLP) in order to learn correlations between the individual views. Each CNN-head is built of five blocks sharing the same structure. Overfitting is prevented by applying a dropout layer [90] after each block. The retention probability is set to $p_{\text{CNN}} = 0.25$. The detailed parameters of the CNN layers are given in Tab. 5.1. The three resulting feature vectors are concatenated and fed into the first fully connected layer `FC1`. After the first two fully connected layers, dropout with retention probability $p_{\text{FC}} = 0.5$ is used. The final fully connected layer `FC3` is followed by a sigmoid function and predicts the desired marginal probabilities of people present in the scene (5.59).

Training End-to-end training of the multi-view CNN is achieved by formulating the estimation of the desired marginal probabilities in (5.59) as a binary classification problem, thus using the binary cross-entropy loss

$$l_{\text{bce}} = -\frac{1}{n}\sum_{i=1}^{n} y_i \cdot \log \hat{p}(x_i) + (1 - y_i) \cdot \log\left(1 - \hat{p}(x_i)\right) , \tag{5.60}$$

with $\mathbf{y} = (y_1, \ldots, y_n)^\mathsf{T} \in \{0,1\}^n$ being the ground truth scene configuration and $\hat{p}(x_i)$ being the predicted probability of a person present at cell u_i. Training samples are created by drawing a random scene configuration,



(a) (b) (c)

(d) (e) (f)

Figure 5.11 Synthetic depth images generated from the scene model for one specific scene configuration $\mathbf{x}$ in sensor view one. (a) shows the synthetic depth image based on the static person model as introduced in Sect. 4.2.1. (b)-(f) show four independently drawn samples from the proposed randomized person model for the scene configuration $\mathbf{x}$.

used as ground truth $\mathbf{y}$ and as input to generate randomized synthetic depth images in perspective of each sensor, which are used as input for the multi-view CNN. An overview of the end-to-end training process is depicted in Fig. 5.10. For training, back propagation and mini-batch gradient decent is used. As optimizer, we use Adam [56].

5.3.2 Randomized Synthetic Depth Image Generation

The randomized generative scene model is an extension of the generative model proposed in Sect. 4.2.1. 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, cf. Fig. 5.11(a). Synthetic depth image generation is obtained by placing 3D person models in the scene depending on the provided scene configuration $\mathbf{x}$. Each 3D person model is rendered into the perspective of each sensor $\mathcal{S}_c$ using the given extrinsic and intrinsic camera parameters. We extend the static person model by introducing a parameterized 3D person model to express different shapes of persons in the scene. Randomization is achieved by treating the parameters of the person model as random variables. Each individual person model is defined by a set of vertices $\mathcal{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_m\}$ with $\mathbf{v} = (x, y, z)^\intercal$ and a set of faces $\mathcal{F}$ where each face is given by a triple of vertices. The set of vertices $\mathcal{V}$ is split into two disjoint subsets $\mathcal{V} = \mathcal{V}_{\text{cyl}} \cup \mathcal{V}_{\text{sph}}$, where the vertices $\mathcal{V}_{\text{cyl}}$ correspond to the cylinder (body) and the vertices $\mathcal{V}_{\text{sph}}$ to the sphere (head) of a person model. The separation of the two geometric primitives is used to apply transformations independently on the cylinder or the sphere of a person model. 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 $z = 0$. Diversity in pose and shape is provided by 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. These variants are expressed by the parameterized transformation

$$f(\mathbf{v}; s_x, s_y, s_z, \gamma) = \mathbf{R}(\gamma) \cdot \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & s_z \end{bmatrix} \cdot \mathbf{v}\,, \tag{5.61}$$

which applies non-uniform scaling followed by a rotation around the z-axis with angle γ on a single vertex. To generate a single instance we apply the transformation $f(\mathbf{v}; S_x^c, S_y^c, S_z^c, \gamma)$ to all vertices in set $\mathcal{V}_{\text{cyl}}$ and $f(\mathbf{v}; 1, 1, S_z^s, 0)$ to all vertices in the set $\mathcal{V}_{\text{sph}}$ respectively. The parameters $S_x^c, S_y^c, S_z^c, S_z^s, \gamma$ are considered to be uniformly distributed random variables (cf. Algorithm 3). To get more variations in shape we add independent Gaussian noise to the x, y, z-components of each vertex $\mathbf{v} \in \mathcal{V}$. A detailed description of the sampling process and the assumed parameter distributions are given in Algorithm 3. Fig. 5.11(b)-5.11(e) show exemplary sampled synthetic depth images for a scene configuration $\mathbf{x}$.

Algorithm 3 Randomized generation of synthetic depth images.

```
procedure SAMPLEFROMGENERATIVEMODEL
    $\mathcal{V}_{\text{cyl}}, \mathcal{V}_{\text{sph}}, \mathcal{F} \leftarrow$ init()    ▷ init with default model
    $h \sim \mathcal{U}(2, 6)$    ▷ drawn number of expected persons
    $\mathbf{x} \sim \mathcal{B}(1/h)$    ▷ draw scene configuration
    for all $x_i = 1$ do    ▷ iterate over cells with a person
        $S_x^c, S_y^c \sim \mathcal{U}(0.5, 1.5)$
        $S_z^c \sim \mathcal{U}(0.85, 1.15)$
        $S_z^s \sim \mathcal{U}(0.85, 1.15)$
        $\gamma \sim \mathcal{U}(0, \pi)$
        ${\mathcal{V}'}_{\text{cyl}} \leftarrow \{f(\mathbf{v}; S_x^c, S_y^c, S_z^c, \gamma) | \mathbf{v} \in \mathcal{V}_{\text{cyl}}\}$
        ${\mathcal{V}'}_{\text{sph}} \leftarrow \{f(\mathbf{v}; 1, 1, S_z^s, 0) | \mathbf{v} \in \mathcal{V}_{\text{sph}}\}$
        $t_x, t_y \sim \mathcal{U}(0, 0.1)$    ▷ draw position offset
        for all $\mathbf{v} \in {\mathcal{V}'}_{\text{cyl}} \cup {\mathcal{V}'}_{\text{sph}}$ do
            $\zeta_x, \zeta_y, \zeta_z \sim \mathcal{N}(0, 0.04)$    ▷ draw AWGN
            $\mathbf{v} \leftarrow \mathbf{v} + (\zeta_x, \zeta_y, \zeta_z)^\intercal$
            $\mathbf{v} \leftarrow \mathbf{v} + (l_{i,x}, l_{i,y}, 0)^\intercal$    ▷ move to grid pos. $l_i$
            $\mathbf{v} \leftarrow \mathbf{v} + (t_x, t_y, 0)^\intercal$    ▷ add position offset
        end for
        renderer.addMesh(${\mathcal{V}'}_{\text{cyl}} \cup {\mathcal{V}'}_{\text{sph}}, \mathcal{F}$)
    end for
    return renderer.getDepthImages()
end procedure
```

6 Evaluation

In this chapter we present quantitative and qualitative results for the inference methods introduced in this thesis. Due to the lack of appropriate data sets for multi-view people detection in depth images, we first introduce a novel data set in Sect. 6.1. In order to quantitatively evaluate and compare the proposed probabilistic inference methods, we apply a threshold to the resulting marginal probabilities to obtain detections and report results in precision-recall space (Sect. 6.2). The experiments focus on the evaluation of the proposed mean-field variational inference methods (Sect. 6.4). We quantitatively compare the mean-field approach as a frame-by-frame detector (Sect. 6.4.1) with state-of-the-art monocular multi-view approaches (listed in Sect. 6.3). Sect. 6.4.2 examines the effect of the temporal context on the mean-field variational inference optimization. In order to put the results into perspective, we additionally compare with the proposed end-to-end multi-view CNN architecture in Sect. 6.5. Since the MAP inference method cannot be directly compared with the other methods due to its dependence on the initialization of the continuous NLLSQ optimization (cf. Sect.5.1), we present qualitative results in Sect. 6.6. Finally, we conclude this chapter with an extensive discussion of the presented results in Sect. 6.7.

6.1 Data Set

To the best of our knowledge, no publicly available data set covers the scenario of top-view people detection using multiple depth sensors with overlapping fields of view. Therefore, we introduce a novel data set to

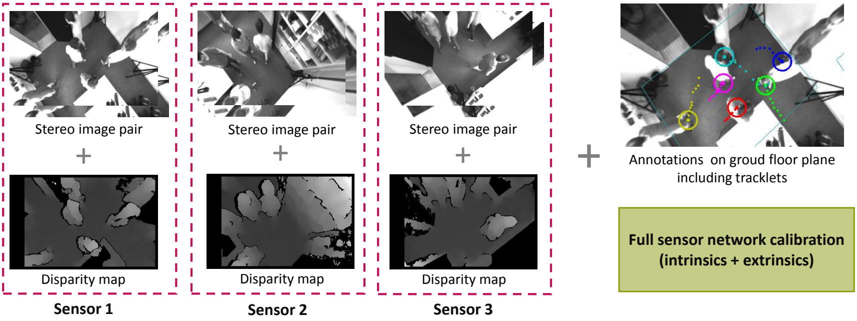


Figure 6.1 Overview of the MULTIPLE data set for top-view indoor people detection. The data set provides raw stereo image pairs as well as disparity maps obtained from three calibrated stereo vision sensors. The presence of individuals is annotated in ground plane coordinates.

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 distributions of individuals in the scene (see Fig. 6.2). The stereo sensors are passive and use block matching to calculate the disparity maps.

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. 4.1). They cover a visible area of approximately $20\,\mathrm{m}^2$ with up to six individual people present in the scene, entering and leaving the visible area multiple times. The data set consists of 2800 annotated multi-view frames[1], captured with a resolution of 376×240 pixel each, providing raw rectified stereo image pairs as well as disparity maps. In total, we annotated the ground level plane locations of more than 12000 targets. Additionally, we associated each detection with a track to allow for full detection and tracking evaluation. The data set is publicly available and

[1] A multi-view frame contains an image from each sensor at one time step. Depending on the context this is also referred to as a temporal frame.

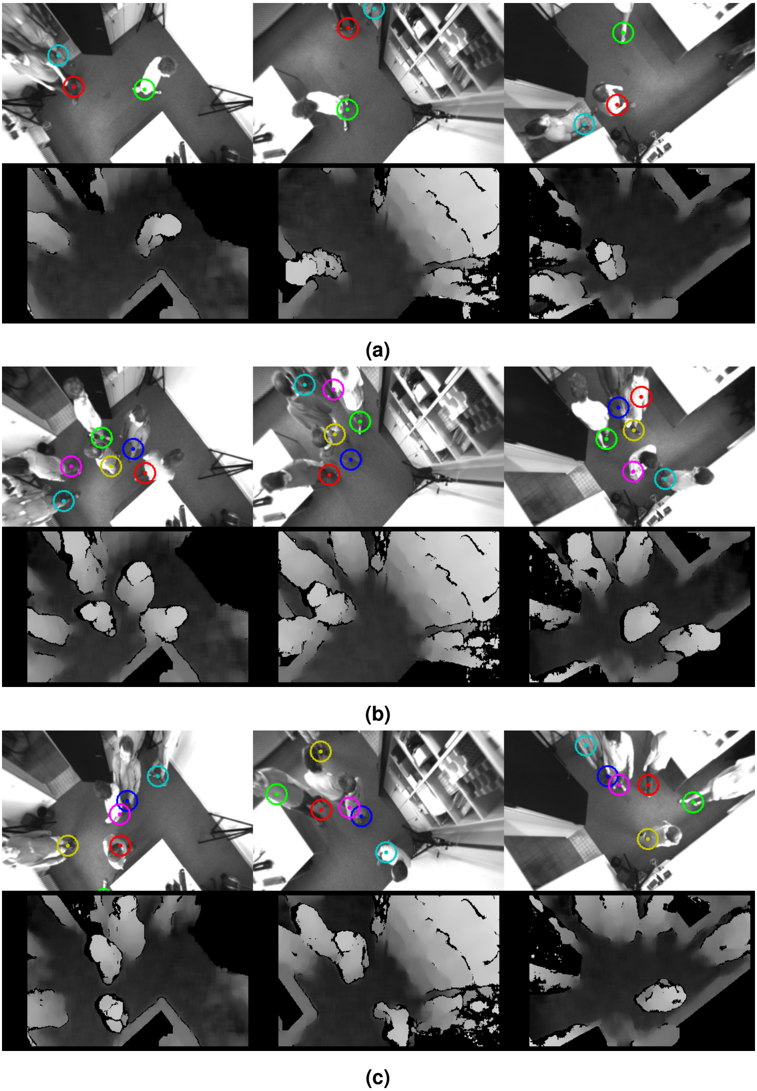
referred to as *MULTIPLE (Multi-View Intensity-Depth Data Set for Top-View Indoor People Detection)*[2].

6.2 Metrics

In order to obtain a set of detections, we threshold the output map of marginal probabilities. Let $\rho \in [0, 1]$ be the detection threshold. By applying the threshold to the inference results in the form of marginal probabilities $q(\mathbf{x})$, the set of detections for one particular value of ρ is given as the grid cell locations $\{u_i \mid q_i(x_i) > \rho\}$. For the quantitative evaluation the resulting detections are matched to the ground truth data by a nearest-neighbor search. A detection is considered to be a true positive if it is within a radius of $30\,cm$ of the ground truth (measured on the ground plane in 2D world coordinates).

In contrast to typical object detection methods, the resulting detections are given in ground plane world coordinates, rather than bounding boxes in images coordinates. Therefore, classical object detection metrics such as the *(Generalized) Intersection Over Union* (IoU) [84] are not suitable for our setup. A typical metric for reporting detection results are the *Receiver Operating Characteristic* (ROC) curves. However, if the underlying classification problem is imbalanced, the ROC metric does not reflect the true performance of a classification algorithm. As an alternative to the ROC metric, the precision-recall space provides a metric for evaluating the classification performance on highly class imbalanced tasks (cf. [87]). For a comprehensive discussion on the relationship between ROC and precision-recall curves we refer to Davis *et al.* [29]. Considering the present detection problem as binary classification reveals that the problem is highly class imbalanced, since only a few cells are actually occupied by an individual. For the quantitative evaluation of the proposed methods we therefore use the precision-recall space.

[2] Subsets of MULTIPLE are introduced in [109, 111]. The full data set is available at https://www.h-ka.de/isrg/publications/multiple.



(a)

(b)

(c)

Figure 6.2 Exemplary samples from the MULTIPLE data set. The first row of each sample shows the rectified gray scale image with ground truth annotations on the ground plane, where each individual is marked with a unique color. The second row shows the raw depth images.

Let TP, FP, FN are the counts of the true positives, false positives and false negatives, respectively. The precision is then defined as the ratio

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}\,, \tag{6.1}$$

where the denominator TP + FP is given as the total count of detections obtained by the evaluated method. Thus, in the context of object detection, the precision can be interpreted as the proportion of correct detections. However, the precision does not account for the false negatives, in this context the individuals present in the scene not detected. To express the trade-off between false positives and false negatives, the precision is considered with respect to the recall, given as the ratio

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}\,. \tag{6.2}$$

In the context of people detection the denominator TP + FN can be identified with the number of individuals present in the ground truth. Hence, the recall can also be identified as sensitivity or true positive rate.

To report the performance of our methods we provide precision-recall curves (e. g. Fig. 6.3) as a function of the detection threshold ρ. To summarize the performance in precision-recall space, we additionally report the F1-Score as well as the *area under the curve* (AUC). The F1-Score is defined as the harmonic mean of precision and recall:

$$F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}\,. \tag{6.3}$$

To summarize the precision-recall performance in one figure, we report the best F1-Score achieved by a particular approach.

The AUC (also referred to as average precision) is defined as the (approximative) integral of the precision-recall curve. In practice, the integral is approximated by the discrete sum over $J \in \mathbb{N}$ threshold values

$$\mathrm{AUC} = \sum_{j=0}^{J-1} (\mathrm{recall}(\rho_j) - \mathrm{recall}(\rho_{j+1})) \cdot \mathrm{precision}(\rho_j)\,, \tag{6.4}$$

with $\mathrm{precision}(\cdot)$ and $\mathrm{recall}(\cdot)$ being a function of a particular threshold value $\rho \in \{\rho_0, \ldots, \rho_J\}$ uniformly covering the interval $[0, 1]$.

6.3 Approaches for Comparison

As mentioned in Sect. 2, only a few approaches in literature rely on depth images for multi-view people detection in the top-view. To the best of our knowledge, the only method directly comparable to our set up is the work by Tseng *et al.* [96]. However, their approach hinges on high quality depth data and neither provides a publicly available implementation nor a data set. To put our results into perspective, we therefore compare the methods proposed in this thesis 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. In detail, the methods to be compared are:

- **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 ground plane. The final detections on the ground plane are obtained by proximity clustering.
- **POM** [38] is methodically the most related approach to the proposed probabilistic inference methods. However, it operates on binary input observations only (cf. Sect. 3.2 for a comprehensive discussion of POM). In the original paper, the binary foreground segmentation

masks are obtained by monocular video cameras. In contrast, we use the same depth based foreground segmentation masks as in our approach for a fair comparison. Furthermore, the grid layout and the camera calibrations are identical to our setup for better comparability. For the experiments we use the original C++ implementation[3] provided by the POM authors.

- **Deep Occlusion** [10] is a current state-of-the art end-to-end architecture for multi-view person detection (a more detailed description is given in Sect. 2.2.1). As input, we stack the given gray scale observations to a three channel image to be compatible with the RGB architecture. Due to the lack of a large data set, we use the available pre-trained model[4] without any further supervision. Considering that the model is pre-trained with RGB images, using stacked gray scale images for inference may have a negative impact on the prediction performance.

6.4 Probabilistic Multi-View People Detection in Discrete Latent Space

In this section we qualitatively and quantitatively evaluate the main contribution of the present thesis: The performance of the probabilistic people detection in discrete latent space by mean-field variational-inference (Sect. 5.2). Since we compare with methods for frame-by-frame detection, we first report comparative results of our mean-field method, omitting the temporal context (Sect. 6.4.1). In Sect. 6.4.2 we show the impact of additionally using the temporal context on the detection performance.

[3] https://www.epfl.ch/labs/cvlab/software/tracking-and-modelling-people/pom/ (accessed 30.04.2021)

[4] https://github.com/pierrebaque/DeepOcclusion (accessed 30.04.2021)

Table 6.1 Performance of the evaluated approaches (without temporal context).

	AUC	F1-Score	Precision	Recall
Ours (MF-VI Data Term)	**0.98**	**0.95**	**0.96**	**0.95**
POM [38]	0.93	0.93	0.91	**0.95**
DoG-Detector	0.76	0.86	0.93	0.79
Deep Occlusion [10]	0.58	0.70	0.80	0.62

For all experiments based on the discrete latent space we employ a ground plane grid with 15×12 grid points, corresponding to a horizontal and vertical distance of $33\,\mathrm{cm}$ between adjacent grid points.

6.4.1 People Detection for One Time Step

The results in this section are based on the experiments previously published in [108]. We compare our mean-field variational inference approach applied to one multi-view frame, omitting the temporal context as introduced in Sect. 5.2.1.

During our experiments we have noticed that our approach is quite sensitive to the initial marginal probabilities $q_i^{\text{init}}(x_i)$. If the initial occupancy probability is too small, the expectation in (5.23) 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 $q_i^{\text{init}}(x_i) = p(x_i) = 0.5$. The design parameter of the asymmetric image similarity $\delta_{\text{asym}}(\cdot, \cdot)$ (5.26) is set to $\alpha = 1.25$, to penalize unexplained observations. The standard deviation of the measurement noise σ_{obs} is set to a default value of $2\,\mathrm{cm}$.

6.4.1.1 Quantitative Results

For the quantitative evaluation we use a subset of the MULTIPLE data set of 2200 consecutive multi-view frames[5]. Fig. 6.3 depicts the performance

[5] Only every 10th consecutive multi-view frame is used to calculate the precision-recall metric.

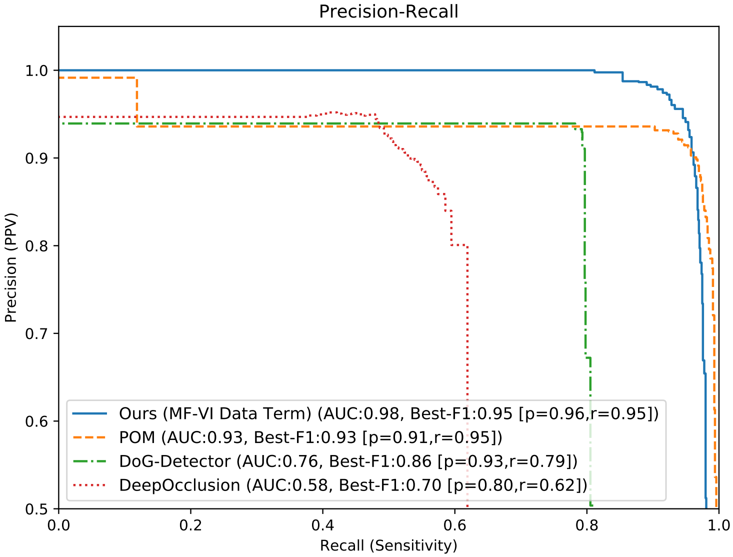


Figure 6.3 Precision-recall curves showing the performance (precision range $[0.5, 1]$) of the mean-field variational inference approach without temporal context, over all views.

of the examined approaches over all views. The results show, that the data set MULTIPLE is very challenging for deep learning architectures, such as Deep Occlusion [10], without domain specific fine-tuning. There are two aspects to put the performance of Deep Occlusion into perspective: (i) due to the vertical top-view, the appearances of people are drastically different compared to the classical profile view; (ii) the original architecture was trained on RGB images rather than gray scale images.

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 [38] operates on binary input images it achieves remarkable performance in our setting. However, our approach outperforms POM in terms of precision, resulting in a better area under the curve value

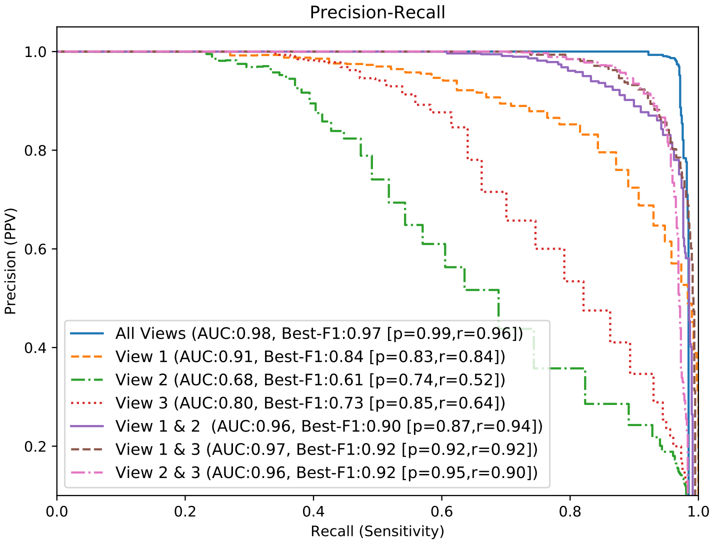


Figure 6.4 Precision-recall curves for different combinations of views. For a fair comparison only the people visible in all three views are taken into account.

(AUC) as well as in a better F1-Score (see Table. 6.1). While POM is quite sensitive (every structure of significant size is detected as an individual person), our approach is more restrictive due to a more expressive forward model, leading to an increased precision. Our hypothesis is that this effect might be even stronger in more complex real-life scenarios, due to the following reasons: (i) foreground images might include additional objects, such as shopping carts or trolley bags; (ii) varying illumination conditions yield a higher level of measurement noise. In both cases, the foreground images get more cluttered, demanding a robust people detection method.

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. 6.2 for the fields of view

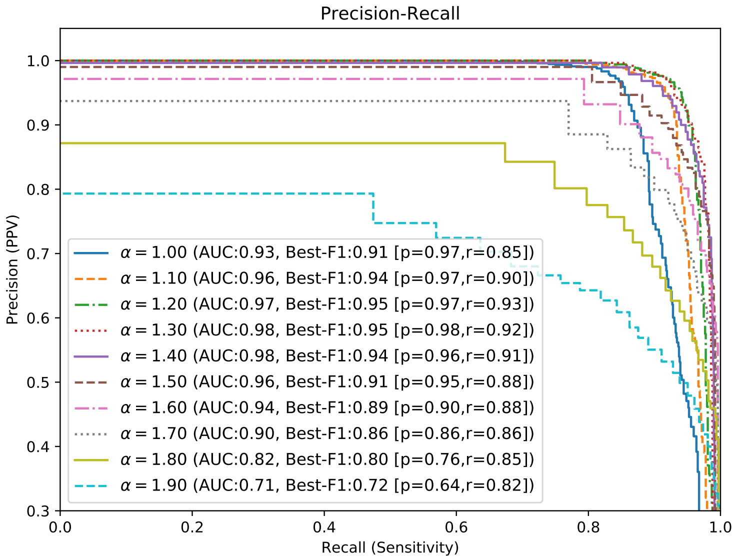


Figure 6.5 Precision-recall curves for different values of the asymmetric image similarity parameter α.

of the sensors). Fig. 6.4 shows the increase in detection performance by leveraging the multi-view information. 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 the best F1-Score of 0.92. Similar results are reported for the combination of View 1 and View 2 with an F1-Score of 0.9 and View 1 and View 3 with an F1-Score of 0.92. Even though View 1 achieves comparably good performance due to the general viewpoint, using the image evidence from all three sensors clearly outperforms all other view combinations. These results clearly demonstrate the capability of the proposed method to jointly leverage the multi-view information in the overlapping image regions.

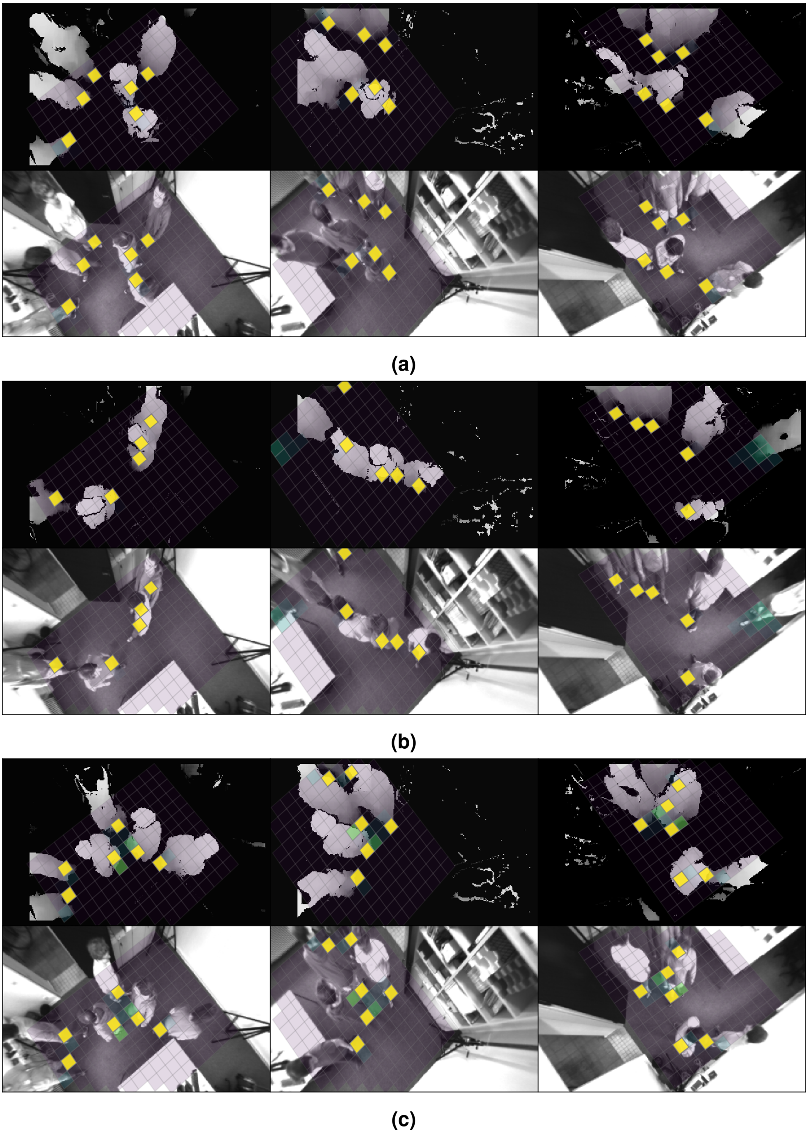
Fig. 6.5 depicts the impact of the asymmetric image similarity parameter $\alpha \in [1, 2]$, introduced in (5.26), on the precision-recall performance. For

$\alpha = 1$ the image similarity function is equal to the L1-norm, while for $\alpha > 1$ the observed pixels not explained by the rendering of the generative scene model are penalized proportional to α. Consequently, increasing α leads to better recall values while potentially decreasing the precision. In this evaluation, a decrease in precision can be observed for $\alpha > 1.4$. Comparing the performance of the symmetric image similarity (i.e. $\alpha = 1$) with the performance of the asymmetric image similarity for $\alpha = 1.3$, a significant increase in performance can be observed, evidenced by the best F1-Score of 0.91 compared to 0.95, respectively.

6.4.1.2 Qualitative Results

Fig. 6.6 and 6.8 show exemplary mean-field optimization results after five iterations. The final marginal probability map is projected onto the ground plane, where purple corresponds to a probability of zero and yellow to one respectively.

Fig. 6.6 highlights some exemplary flawless 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. 6.6(b) gives a particular example for the ability to handle occlusion by exploiting the image evidence from multiple sensors. While the fifth individual is fully occluded on the first view (left image in Fig. 6.6(b)), the second view and partially also the third view include enough image evidence of the fifth individual to predict a probability of occupancy close to one. Fig. 6.6(c) shows that the resulting marginal distributions also include some uncertainty around the peaks of the distribution. This happens in general if the generative scene model cannot precisely explain the observations, e. g. due to discretization errors, measurement noise or the simplified 3D person model (cf. Sect. 4.2.1). However, for the majority of samples the resulting marginal distribution of people present in the scene contains clear peaks with a high confidence.



(a)

(b)

(c)

Figure 6.6 Exemplary mean-field optimization results $\hat{q}(\mathbf{x})$ (without temporal context). (a) and (b) show an estimation of the marginal probability distribution $q(\mathbf{x})$ with clear peaks at grid locations occupied by a person. (c) includes some uncertainty near the peaks of the distribution.

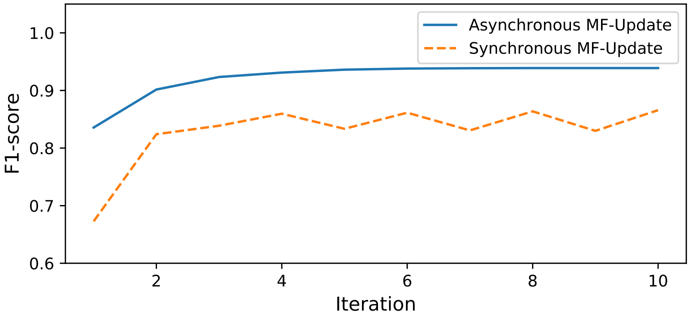
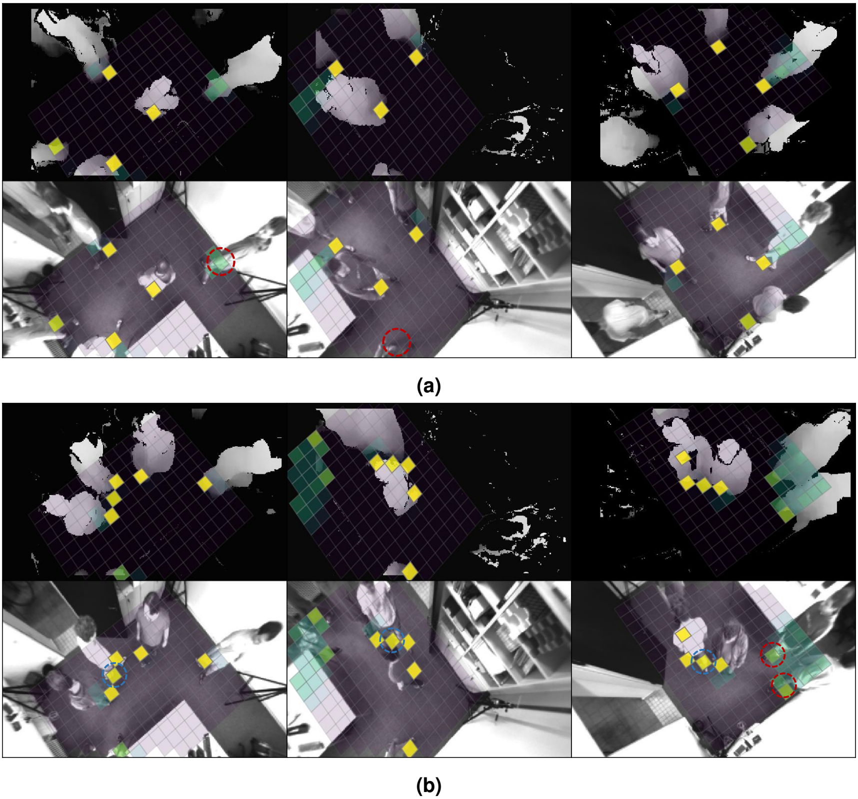


Figure 6.7 Comparison of asynchronous and synchronous mean-field updates considering the best achieved F1-score in each mean-field optimization iteration.

In Fig. 6.8 two mean-field optimization results, including some exemplary faulty detections are shown. Fig. 6.8(a) shows a typical false negative case on the image border, which is the dominant error class occurring in the data set. Due to the stereo-vision 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 in future work, a richer probabilistic sensor model which takes systematically varying noise into account could be employed. Two similar false negative cases can also be observed in Fig. 6.8(b) in the third view. Apart from that, Fig. 6.8(b) includes a false positive detection, indicated by a blue dashed circle. This type of false positive potentially occurs between two individuals due to the discretization error and the simplified 3D person model. However, regarding the present data set, false positives are the minority errors occurring in the evaluation.

In Fig. 6.9, the iterative mean-field optimization process is illustrated for one exemplary frame, for both the asynchronous and the synchronous update strategy (cf. Sect. 3.1.4). In Fig. 6.9(b) the asynchronous mean-field update optimization is illustrated. One can observe that after the first optimization iteration the probability mass is already quite concentrated around the grid locations occupied by a person. In iterations 2 and 3 the probability mass gets more concentrated, leading to a marginal distribu-

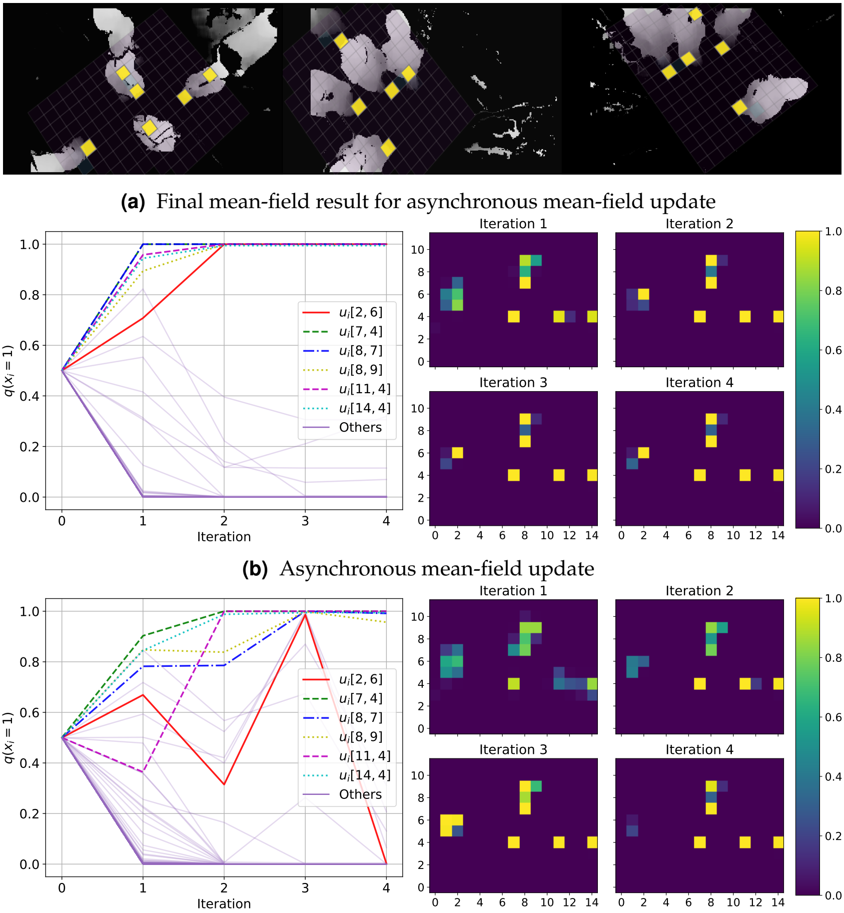


(a)

(b)

Figure 6.8 Exemplary mean-field optimization result $\hat{q}(\mathbf{x})$, including faulty detection results. (a) includes a typical false negative, marked with a red dashed circle. (b) includes a false positive, marked with a blue dashed circle as well two false negatives.

tion with clear peaks. In contrast, Fig. 6.9(c) depicts the optimization for synchronous mean-field updates. The example illustrates typical problems of synchronous mean-field updates. In contrast to the asynchronous update, the probability mass is distributed on more grid cells after the first iteration. Also in the upcoming iterations the simultaneous optimization suffers from oscillating marginal probabilities, evidenced by the grid location $u_i[2, 6]$, shown as red curve in Fig. 6.9(c). Considering the overall best



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

(b) Asynchronous mean-field update

(c) Synchronous mean-field update

Figure 6.9 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.

Table 6.2 Performance of temporal mean-field approximations.

	AUC	F1-score	Precision	Recall
MF-VI Temporal Smoothing	0.94	**0.94**	**0.96**	**0.92**
MF-VI Bayesian Filtering	**0.96**	0.93	0.95	0.90
MF-VI Data Term	0.95	0.91	0.92	0.90
DoG-Detector	0.85	0.83	0.77	0.92

F1-score per iteration, Fig. 6.7 shows that the asynchronous update strategy clearly outperforms the synchronous update strategy. While asynchronous mean-field optimization converges after only few iterations, synchronous mean-field update suffers from the oscillating effects mentioned above. In the remainder of this evaluation, the asynchronous mean-field update strategy is used.

6.4.2 People Detection with Temporal Context

The results in the previous section are based on typical frame-by-frame detection to be comparable to state-of-the art approaches. In contrast, in this section we will focus on the evaluation of the additional impact of exploiting the temporal context. Hence, we will evaluate the mean-field variational inference approximations for (i) the Bayesian filtering distribution $p(\mathbf{x}_t \mid \mathbf{o}_{1:T})$ as proposed in Sect. 5.2.2; (ii) the full posterior distribution $p(\mathbf{x}_{1:T} \mid \mathbf{o}_{1:T})$ (referred to as temporal smoothing) as introduced in Sect. 5.2.3. This section is based on the results previously published in [109].

6.4.2.1 Quantitative Results

In Sect. 6.4.1 we report quite strong detection performance for the mean-field approach omitting the temporal context (MV-VI Data Term), evidenced by the best F1-score of 0.95 and an AUC of 0.98. Since these results are nearly reaching the optimum, we evaluate the exploitation of the tem-

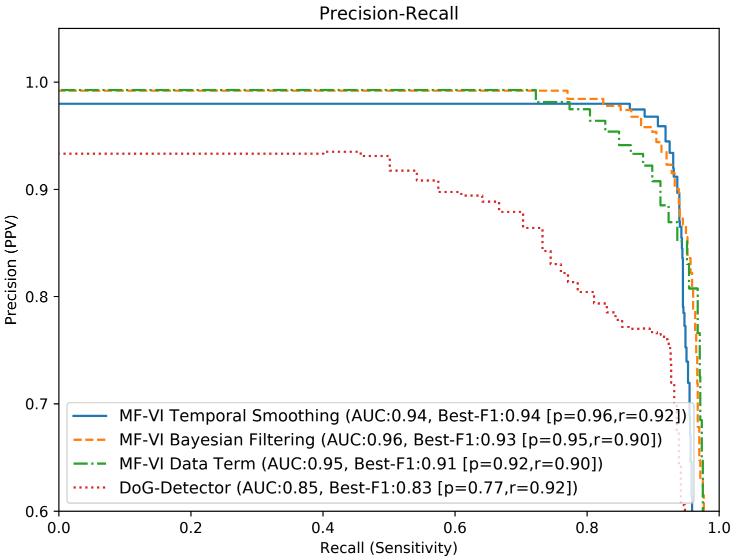


Figure 6.10 Precision-recall curves showing the performance of our approach with and without temporal context.

poral context on a slightly more challenging[6] disjoint image sequence. The sequence consists of 300 consecutive multi-view frames, which are a sub sequence of the MULTIPLE data set introduced in Sect. 6.1. As a consequence of the more challenging sequence the performance of the MV-VI Data Term drops to best F1-score of 0.91 and an AUC of 0.95 (cf. Fig. 6.10). Subsequently, we analyze the effects of using the temporal context by comparison of the following methods:

- **MF-VI Temporal Smoothing** refers to the approximation of the full posterior distribution $p(\mathbf{x}_{1:T} \mid \mathbf{o}_{1:T})$ as introduced in Sect. 5.2.3. During evaluation, we observed that the influence of the proposed future term in (5.53) on the quantitative results is negligible for the proposed update strategy and motion model. Therefore, we set

[6] Compared to the sequence used for the single time step case, the sequence contains slightly more individuals present at the image border, making people detection more challenging.

the weighting parameters to $\lambda_{\text{past}} = 0.65$, $\lambda_{\text{future}} = 0.0$. In consequence the presented results are based only on the past and data term defined in (5.46), (5.50) respectively. Assuming only moderate movement of individuals between two consecutive frames, the parameter μ_{self}[7] in the dynamics model (4.40) is set to $\mu_{\text{self}} = 0.8$. The normalization weights are set according to (4.42). For evaluation, we run six mean-field iterations.

- **MF-VI Bayesian Filtering** refers to the approximation of the filtering distribution $p(\mathbf{x}_t \mid \mathbf{o}_{1:T})$ as proposed in Sect. 5.2.2. The dynamics model parameter is set to $\mu_{\text{self}} = 0.2$. As for the temporal smoothing method we perform six mean-field iterations for evaluation.

- **MF-VI Data Term** refers to the approximation of the posterior distribution $p(\mathbf{x}_t \mid \mathbf{o}_t)$ limited to the observations from a single time step and a uniform prior, as introduced in Sect. 5.2.1 and evaluated extensively in Sect. 6.4.1.

- **DoG-Detector** To put the results into perspective we compare them with the Difference of Gaussian blob detector as introduced in Sect. 6.3.

Fig. 6.10 shows a comparison of the precision-recall performance for the methods mentioned above. Although the performance of the mean-field variational inference approach without temporal context (MF-VI Data Term) is already quite high (best F1-score of 0.91), the results show that the exploitation of the temporal context can increase the overall precision and recall performance, evidenced by a maximum F1-score of 0.94 for the MF-VI Temporal Smoothing approach and 0.93 for the MF-VI Bayesian filtering approach, respectively. Considering that the temporal smoothing approach is a batch processing method, the full sequence of observations is optimized jointly. In contrast, the Bayesian filtering distribution is defined recursively, where the past is condensed in the mean-field approximation

[7] Reflecting that a person will stay on the current location with a probability of μ_{self}.

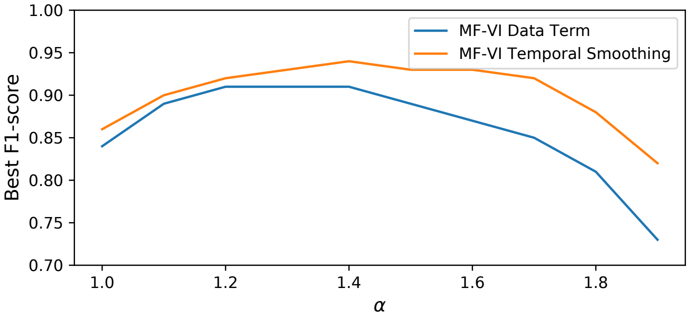


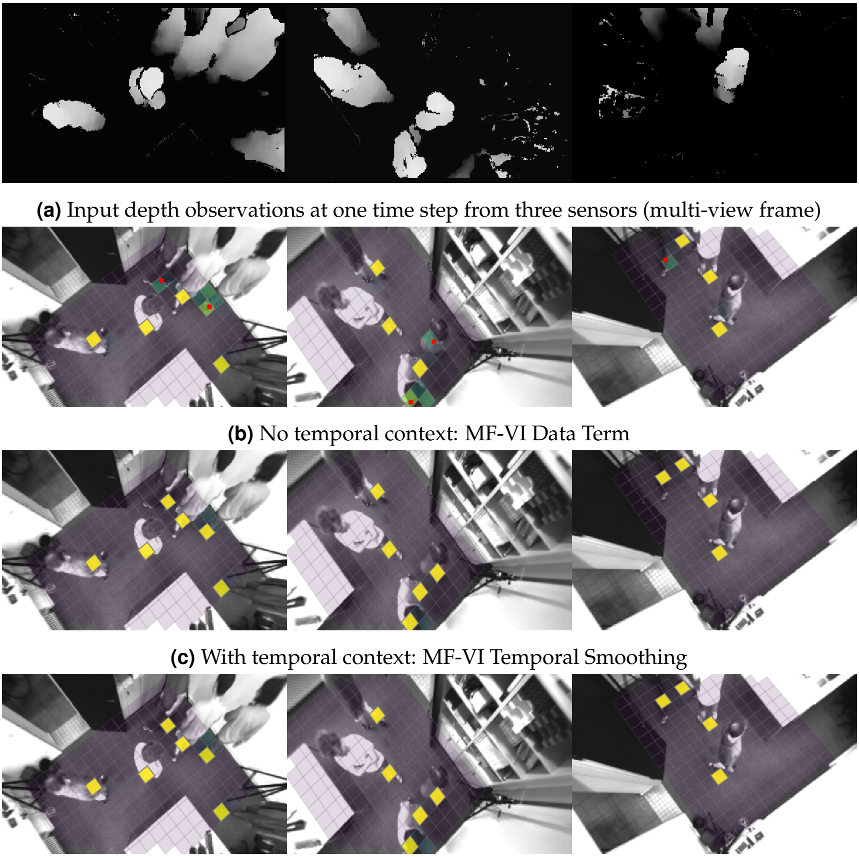
Figure 6.11 Dependence on hyperparameter α evaluated for MF-VI Data Term and MF-VI Temporal Smoothing.

of the previous time step $\hat{q}(\mathbf{x}_{t-1})$. In practice, this enables to use the Bayesian filtering approximation as a classical frame-by-frame detector in real-time applications. From this perspective, it is remarkable that the Bayesian filtering method does exhibit a similar performance on the given data set compared to the temporal smoothing approach. Still, the results must be interpreted carefully, since the current data set is limited and the mean-field approximation without usage of any temporal context already performs quite well. Our hypothesis is that in order to capture the full capabilities of the temporal approaches, evaluation on a more challenging large-scale data set is required.

In Fig. 6.11, the best F1-score depending on α is plotted. The results indicate that the temporal smoothing is slightly less sensitive to the choice of the image similarity weighting parameter α, compared to the MF-VI Data Term approach. The temporal regularization might be able to compensate for the suboptimal single frame detections.

6.4.2.2 Qualitative Results

Fig. 6.12 and Fig. 6.13 illustrate exemplary comparative mean-field optimization results with and without using the temporal context. In Fig. 6.12 an exemplary multi-view frame leveraging the temporal regularization is



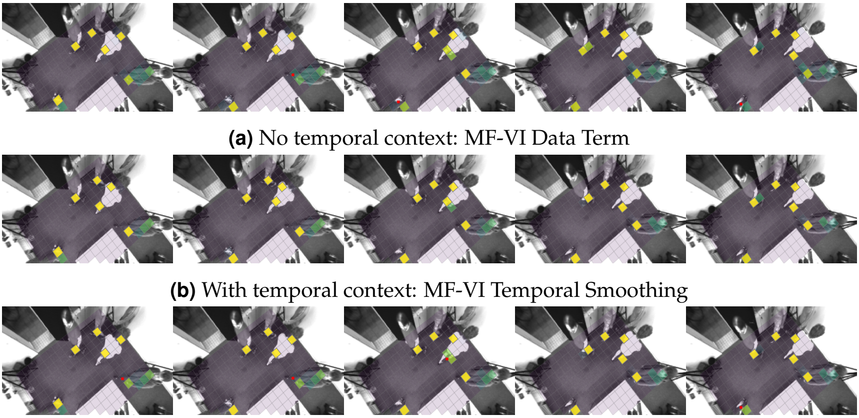
(a) Input depth observations at one time step from three sensors (multi-view frame)

(b) No temporal context: MF-VI Data Term

(c) With temporal context: MF-VI Temporal Smoothing

(d) With temporal context: MF-VI Bayesian Filtering

Figure 6.12 Exemplary mean-field optimization results depicted for one multi-view frame (a). (b,c,d) show the resulting marginal probability map projected onto the ground plane. False negatives are marked with a red dot.



(a) No temporal context: MF-VI Data Term

(b) With temporal context: MF-VI Temporal Smoothing

(c) With temporal context: MF-VI Bayesian Filtering

Figure 6.13 Mean-field results for five consecutive frames, projected into sensor view one.

shown. Without the temporal context (Fig. 6.12(b)), the estimated marginal probability distribution contains high uncertainty near the two targets due to partial visibility and measurement noise, leading to a false negative detection in both cases. Exploiting the temporal context can resolve those uncertainties, leading to a marginal distribution with clean peaks for both, the MF-VI Temporal Smoothing (Fig. 6.12(c)) and the MF-VI Bayesian Filtering approach (Fig. 6.12(d)).

Similar effects can be observed in Fig. 6.13, where results for an exemplary sequence of five consecutive frames (shown only in sensor view one) are depicted. Without exploiting the temporal context (Fig. 6.13(a)), detection of the person at the bottom right-hand side is very unstable over time. In contrast, by applying the temporal smoothing these predictions are getting stabilized as evidenced by Fig. 6.13(b). In this sequence, the MF-VI Bayesian Filtering approach does not perform quite as well as the MF-VI Temporal Smoothing method. Studying the first two frames, the filtering approach does not improve the detection results compared to the MF-VI Data Term (Fig. 6.13(a)), while the temporal smoothing clearly does.

However, once there is some significant probability mass predicted by the data term (as observed in Fig. 6.13(a), frame three), the prediction is stable in the upcoming frames four and five as evidenced in Fig. 6.13(c). This reveals a general difference between the temporal smoothing and the temporal filtering approach. While in the temporal smoothing approach, the marginal distributions are approximated jointly over space and time, in the filtering approach the past is condensed in the approximated distribution from the previous time step. Due to the joint optimization, the temporal smoothing approach effectively incorporates temporal information in order to fill in the gap of noisy predictions.

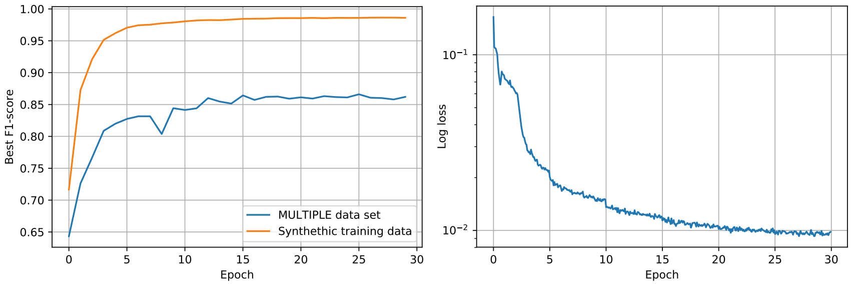
We observed that the run time per frame decreases slightly on average by using the temporal context. This can be explained by the fact that the proposed dynamics model effectively restricts the set of grid cells where a person can be present with a probability greater than zero, thus less mean-field updates need to be evaluated. On a single CPU core[8], our non-optimized Python implementation needs approximately 700 ms per multi-view frame.

6.5 CNN Inference

In this section we evaluate the end-to-end multi-view CNN approach introduced in Sect. 5.3. We investigate whether the proposed black box inference method, trained with synthetic data only, achieves comparable results to the more involved mean-field variational inference methods. In comparison to the probabilistic approaches, the proposed CNN architecture is trained with images obtained by the modified[9] generative scene model (cf. Sect. 4.2.1). Thereby, the same discrete ground plane grid is used to encode the prediction results. Since the proposed CNN operates on a multi-view frame without considering the temporal context, we com-

[8] Intel Core-i7 with 2.9 GHz

[9] As described in Sect. 5.3.2, the depth image output of the generative scene model is randomized.



(a) Best F1-score vs epochs

(b) Training loss

Figure 6.14 Exemplary training process of the multi-view CNN architecture with 150.000 synthetic multi-view frames, sampled from the randomized scene model. (a) shows the best F1-score achieved after each epoch on the MULTIPLE data set as well as on a subset of the synthetic training data. (b) shows the corresponding training loss on a logarithmic scale.

pare it to the single frame detection method MF-VI Data Term. For the experiments we use the same sequence of 2200 frames as in the evaluation of the MF-VI Data Term (Sect. 6.4.1). Since we observed only a slight drop in performance when the input depth images are downscaled by a factor of 2, we used subsampled depth images with a resolution of 188×120 pixel as input. This has the advantage that the GPU memory footprint can be reduced significantly, leading to a faster training process. The results in this section are based on the experiments previously published in [112] by the author.

To train the model, we use 150.000 synthetic multi-view frames (consisting of three depth images each). For the randomized scene model we sample each frame accordingly to Algorithm 3. For the plain scene model we use the same scene configuration sampling strategy, but omit the manipulation of the vertices. We train the model with a batch size of 96 and use the Adam [56] optimizer with a learning rate of 0.001. An exemplary training process is illustrated in Fig. 6.14.

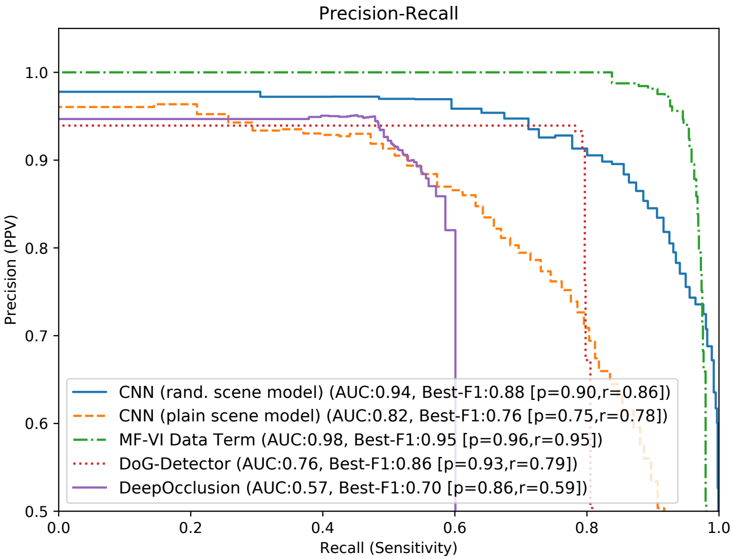
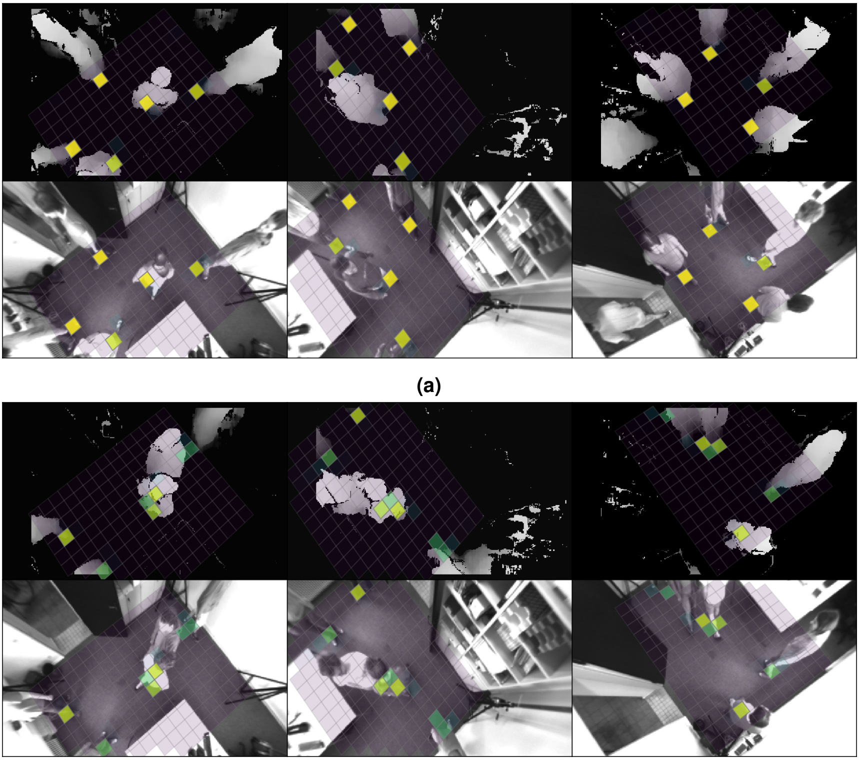


Figure 6.15 Precision-Recall curves showing the performance of the multi-view CNN approach with and without domain randomization.

6.5.1 Quantitative Results

Fig. 6.15 shows the precision-recall performance of the evaluated approaches. While the mean-field variational inference approach (MF-VI Data Term) outperforms the other methods, the proposed end-to-end CNN method trained with synthetic randomized depth images (CNN (randomized scene model)) achieves noticeable results with best F1-score of 0.88. To put these results into perspective, one has to consider that the CNN architecture is a black box inference method only trained with synthetic training data and does not incorporate any further scene knowledge or explicit modeling. Also, the MF-VI Data Term includes an iterative optimization method at inference time. In contrast, the multi-view CNN approach is computationally evolved during training, however inference is just a forward pass. Still, the performance of the CNN approach is not competing with the more involved mean-field inference. Comparing the two manifestations of the proposed architecture CNN (plain scene model) and CNN (randomized scene model), shows that randomizing the scene model has a significant impact on the performance (F1-score of 0.76 for plain person model vs. F1-score of 0.88 for randomized person model). This is a noticeable result, since the randomization of the person model as proposed in Sect. 5.3.2 does not include any specific scene knowledge. Still, it leads to a better generalization of the trained model by narrowing the gap between synthetic images and real-world observations, resulting in a significant increase in detection performance.

6.5.2 Qualitative Results

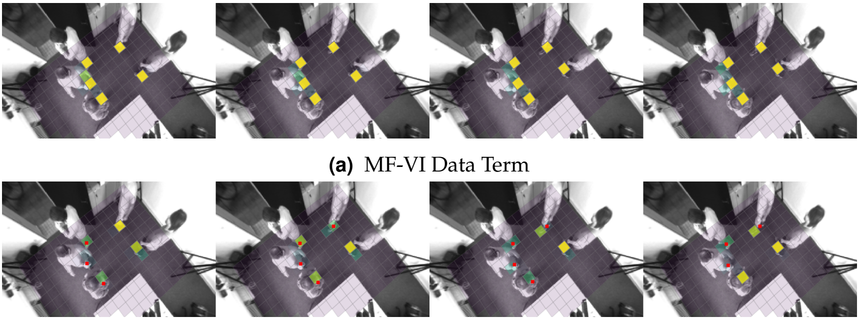


(a)

(b)

Figure 6.16 Exemplary inference results for the multi-view CNN trained on randomized synthetic depth images (CNN (randomized scene model)).

In Fig. 6.16 qualitative results of the proposed CNN architecture trained with randomized synthetic depth images are illustrated for two exemplary multi-view frames. Fig. 6.16(a) shows an exemplary convincing example with clear peaks at grid locations occupied by a person. In contrast, Fig. 6.16(b) shows a particular faulty example. The probability mass is spread around the grid cells occupied by an individual, however the



(a) MF-VI Data Term

(b) CNN (randomized scene model)

Figure 6.17 Results for a sequence of four consecutive multi-view frames (for visualization purpose only sensor view one is depicted). (a) shows the marginal probabilities obtained by mean-field variational inference. (b) depicts the marginal probabilities predicted by the multi-view CNN approach.

marginal probabilities are all quite low. In consequence, none of them clearly predicts the presence of a person, leading to false negative detections. Comparable false negative cases are the dominant error cases in the present data set.

Fig. 6.17 shows the direct comparison of four consecutive frames, projected into sensor view one. This example shows a significant weakness of the proposed CNN end-to-end inference. The results of mean-field variational inference (without using the temporal context) are quite robust over time as evidenced by Fig. 6.17(a). In contrast, the predictions of the CNN architecture are very unstable over the given sequence of consecutive frames. Even the appearance of individuals hardly change in between two consecutive frames, the predicted marginal probabilities significantly differ. For example, the grid cell occupied by the person at the top right-hand side exhibits a high probability in the first frame. However, in the second frame the probability drops significantly, leading to a false negative detection.

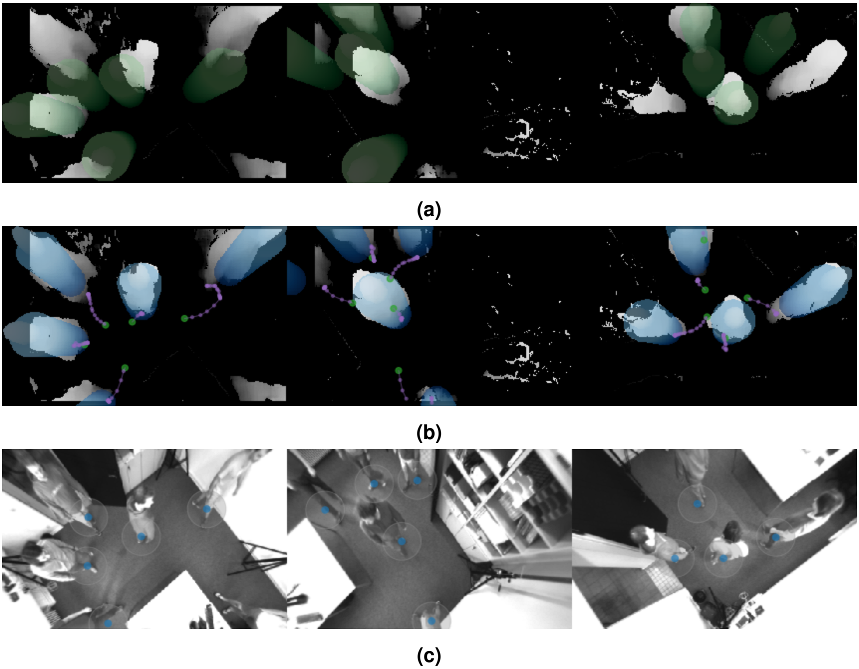
6.6 MAP Inference in Continuous Latent Space

In this section we report results for the MAP inference method introduced in Sect. 5.1.1. A fair comparison of the MAP inference method with the methods evaluated in the previous section is not possible due to the following reasons: (i) in order to formulate the MAP objective, the number of people in the scene has to be known a priori; (ii) the underlying non-linear least squares optimization problem relies on a good initialization. As a consequence, the MAP inference is not suitable as a stand-a-lone method in real-world scenarios, whereas it can be seen as a complementary method to the proposed discrete inference methods, enabling fine-tuned discrete person localization. To illustrate the capabilities of the MAP inference approach, we present qualitative results in this section. In Sect. 6.6.1 we report results obtained by frame-by-frame inference, while in Sect. 6.6.2 we apply the MAP inference method to a sequence of consecutive multi-view frames.

6.6.1 Optimization for One Time Step

In this section, we show MAP optimization results omitting the temporal dynamics as defined in the objective (5.9). The box regularization term is weighted with $\lambda_{\text{box}} = 0.1$ and the pairwise distance regularization term with $\lambda_{\text{dist}} = 1$ respectively. The number of people present in the scene is explicitly given as the ground truth for each multi-view frame. If not reported otherwise, the initial scene configuration $\mathfrak{X} = (\check{\mathbf{x}}_1, \ldots, \check{\mathbf{x}}_m)$ is drawn randomly. It is assumed that each person location $\check{\mathbf{x}}_i \in \mathbb{R}^2$ is drawn independently and distributed uniformly in a rectangular initialization area. By narrowing the initialization area to a reduced edge length of 70% of the rectangular observable area, initial person locations close to the image border are avoided. Fig. 6.18 shows a flawless MAP inference example for one frame of the MULTIPLE data set[10]. Notice that the individual at the

[10] The result is cherry-picked from 10 optimization results, obtained by different randomly drawn initial scene configurations.

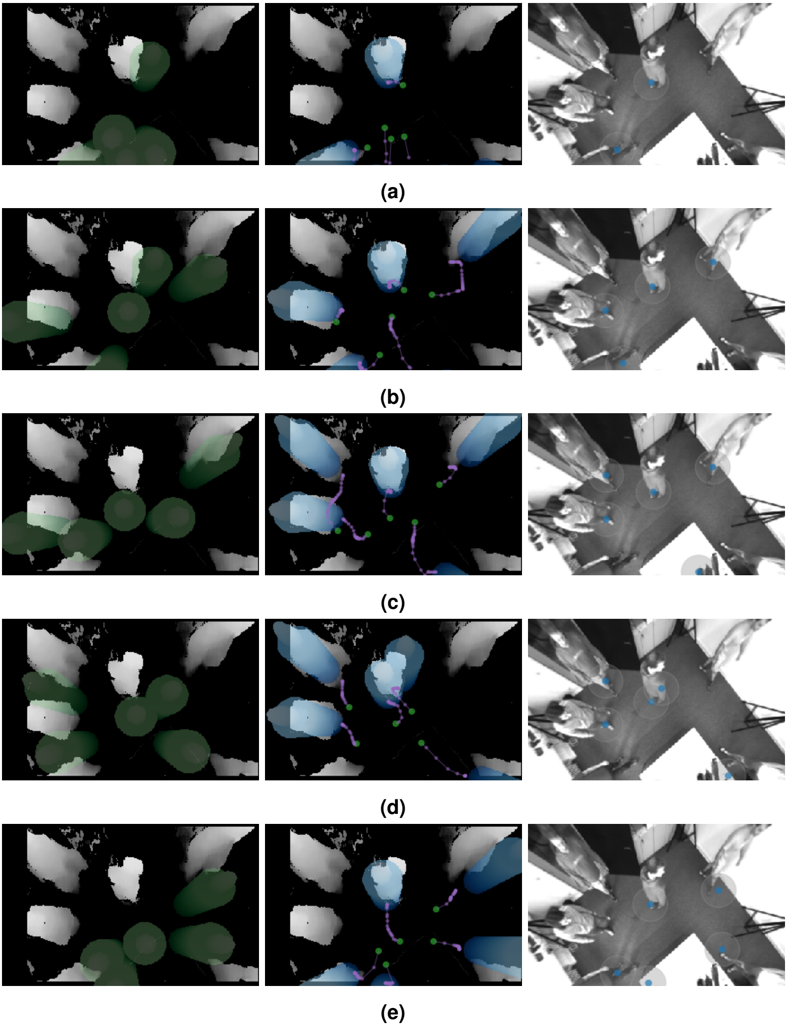

(a)

(b)

(c)

Figure 6.18 Optimization result for the MAP objective without temporal dynamics. (a) shows the depth observations with the synthetic depth images corresponding to the initial scene configuration as green overlay. (b) shows the synthetic depth images corresponding to the final MAP result as blue overlay. The initial positions are given as green dots, the optimization trace is drawn in purple. (c) illustrates the final optimization result projected into the camera view of each sensor.

bottom right of sensor view one is not considered to be part of the ground truth since it is only hardly visible in a single sensor view. In Fig. 6.18(a) one can see that the initial person models in the synthetic depth images are already partially overlapping with the depth observations of the individuals present in the scene. In this case, for each location $\check{\mathbf{x}}_i$ some gradients point in the direction of an individual, which increases the probability of converging to a satisfying local optimum (cf. 5.1.1).



(a)

(b)

(c)

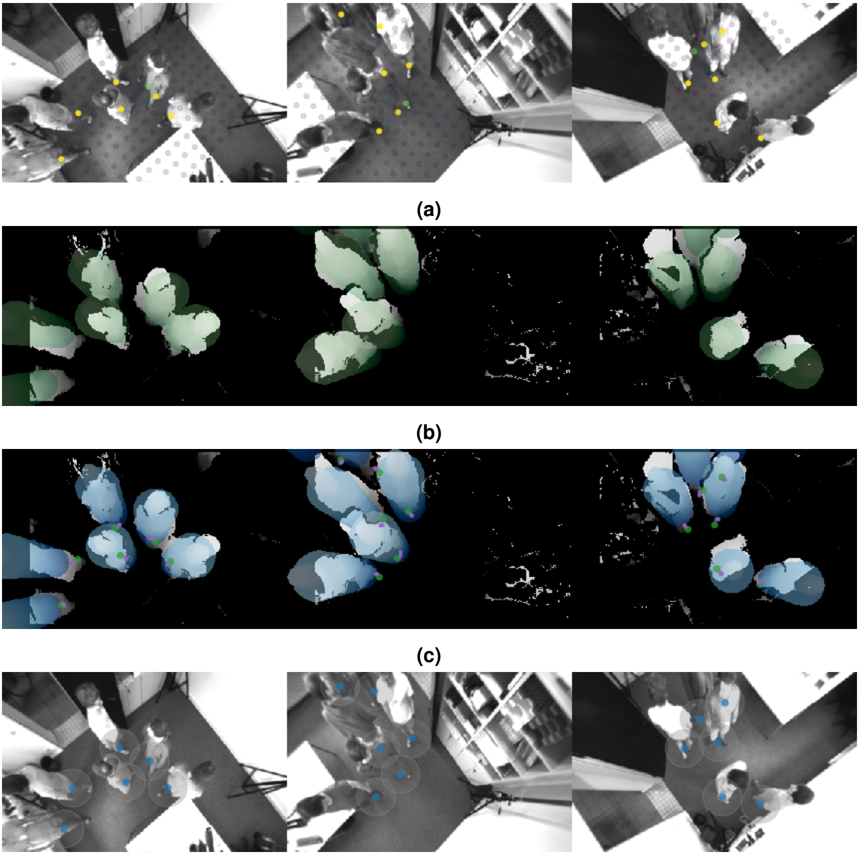
(d)

(e)

Figure 6.19 MAP optimization results for one multi-view frame and five different initial scene configurations (a)-(e). Per scene configuration initialization from left to right: (left) depth observation of sensor one with initial scene model as green overlay ; (middle) final MAP state of scene model as blue overlay and optimization trace in purple; (right) final MAP results projected into the camera view one.

To emphasize the reliance on a good initialization for the proposed continuous optimization method Fig. 6.19 illustrates results for five randomly drawn initial scene configurations. For all independent optimization runs Fig. 6.19 (a)-(e), the given observations can only be partially explained by the model due to a suboptimal initialization. Fig. 6.19(a) shows a particular ill-posed initialization. Four of the five person locations are initialized close to each other at the bottom of the observable area. After optimization one of those person locations successfully explains the individual at the bottom left in the scene. However, the remaining person locations are finally located at the border of the scene. This can be explained as follows: If a rendering of an individual does not share any overlap with the image evidence of an individual, the data term costs will only be reduced if the synthetic person model moves out of the field of view. During optimization affected person locations move towards the image border until the box prior creates an opposite effect. This behavior can be observed in all given samples in Fig. 6.19 to some extent.

In Fig. 6.20 the combination of the MAP inference method with the mean-field variational inference approach introduced in Sect. 5.2 is illustrated. Fig. 6.20(a) shows the MF-VI result for a single multi-view frame. Applying a threshold to the marginal probabilities yields a list of detections, used to initialize the continuous optimization problem. By solving for the MAP state, the initial discrete grid based person locations get fine-tuned, leading to more precise locations on the ground plane. In Fig. 6.20(b) the 3D person models corresponding to the MF-VI initialization do already fit the observations quite well. However, studying the overlay image of the final MAP results in Fig. 6.20(c) reveals that the model fit can be further improved. While these improvements are limited due to the chosen discrete grid with a horizontal and vertical distance of $33\,\mathrm{cm}$ between adjacent grid points (cf. Sect. 6.4), the fine-tuning can be a crucial part for applications demanding a precise localization in image coordinates. Another potential application is to run the MF-VI detector only for every



(a)

(b)

(c)

(d)

Figure 6.20 Exemplary results of MAP inference initialized with the mean-field detection results given in (a). (b) shows the depth observations with the synthetic depth images corresponding to the initial scene configuration as green overlay. (c) shows the synthetic depth images corresponding to the final MAP result as blue overlay, whereas the initial positions are given as green dots and the optimization trace is drawn in purple. (d) illustrates the final optimization result projected into the camera view of each sensor.

k-th frame and apply MAP inference to the gap frames, initialized with the last MF-VI result.

6.6.2 Sequence Optimization

The results in the previous section are based on frame-by-frame inference. In contrast, in this section we apply the full MAP objective given in (5.8) to a sequence of consecutive multi-view frames of the MULTIPLE data set. The additional temporal term is weighted with $\lambda_{\text{temporal}} = 0.01$. To initialize the sequence of scene configurations $\mathfrak{X}_{1:T}$, we first draw one time independent random scene configuration $\mathfrak{X}^{\text{init}} = (\check{\mathbf{x}}_1^{\text{init}}, \ldots, \check{\mathbf{x}}_m^{\text{init}})$ from a uniform distribution as described for the single frame inference in Sect. 6.6.1. Based on the initial scene configuration $\mathfrak{X}^{\text{init}}$, each initial person location $\check{\mathbf{x}}_{i,t}$ for $i \in \{1, \ldots, m\}$ and $t \in \{1, \ldots, T\}$, is independently drawn from the normal distribution $\mathcal{N}\left(\check{\mathbf{x}}_i^{\text{init}}, \sigma^2\mathbf{I}\right)$, with $\sigma = 0.7\,\mathrm{m}$. This ensures that the initial scene configuration is not ill-posed and huge jumps of an individual from one time step to the other are avoided. At the same time the proposed initialization exhibits enough randomness to explore the scene configuration space, in order to increase the probability of successfully converging to an optimal fit. Our experiments showed that by initializing each $\check{\mathbf{x}}_{i,t}$ independently, the temporal regularization costs will dominate the data term costs, eventually leading to an ill-posed optimization problem.

The major strength of the additional temporal regularization is that the optimization becomes more independent of a particular initialization. This effect can be observed in Fig. 6.21, where the optimization process for six consecutive multi-view frames is illustrated. Applying the proposed joint optimization to this sequence leads to a satisfying fit for all time steps. In contrast, solving for the scene configuration for a single time step (while using the same initialization) is not successful for $t = \{3, 4, 5\}$. This is exemplarily evidenced for $t = 4$ in Fig. 6.22. The MAP scene configuration is successfully determined by considering this multi-view frame as part of a sequence of frames, as shown in Fig. 6.22(a). In contrast, using the same

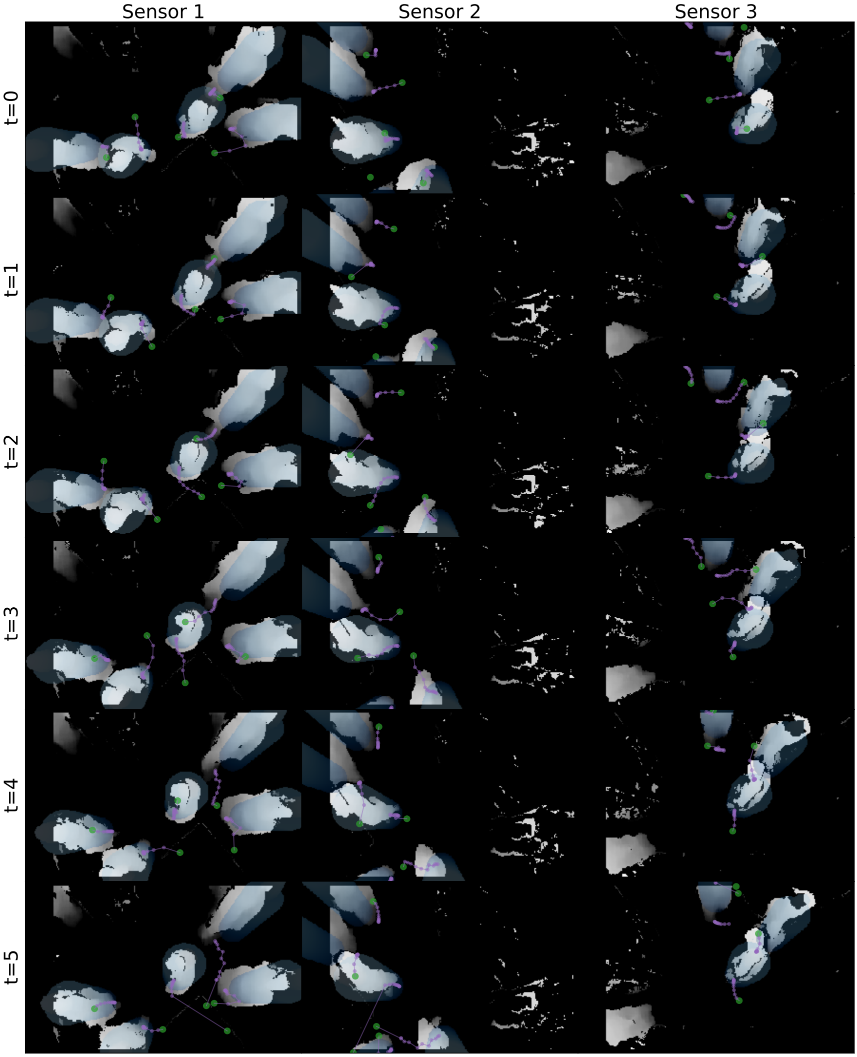
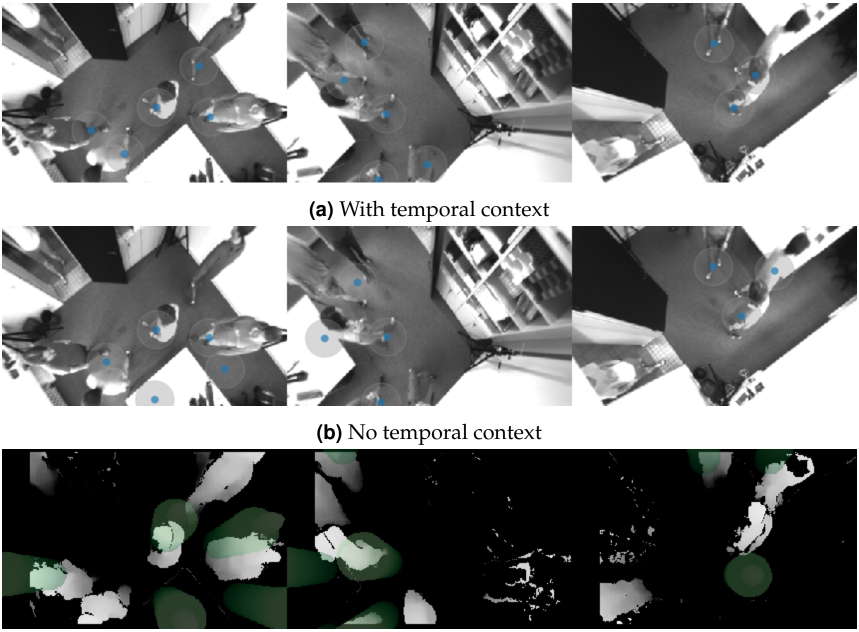


Figure 6.21 MAP optimization result for a sequence of six consecutive multi-view frames. Each row corresponds to one time step. For a explanation of the overlay see description Fig. 6.18.



(a) With temporal context

(b) No temporal context

(c) Initialization

Figure 6.22 Comparison of optimization results for frame at $t = 4$ (cf. Fig. 6.21), with (a) and without temporal context (b). Both optimizations are initialized with the same scene configuration, visualized in (c) as green overlay.

initialization (illustrated in Fig. 6.22(c)) but omitting the temporal context, the optimization process gets stuck in a local minimum far away from the true scene state, see Fig. 6.22(b).

6.7 Discussion

The results reported in Sect. 6.4.1 demonstrate that even without using the temporal context our proposed mean-field variational inference approach achieves strong detection performance, outperforming state-of-the-art monocular multi-view people detection methods. Considering that POM operates only on binary input images, it is remarkable that POM achieves such competitive results. On the one hand this shows the strength of the general idea of generative probabilistic modeling in combination with the mean-field approximation for multi-view people detection. On the other hand our hypothesis is that the difference in performance between POM and our approach would be more significant in more complex real-life scenarios, where the foreground images include additional objects and a higher level of noise. Besides the direct comparison to the literature, the results also indicate that our joint optimization approach effectively leverages the given multi-view information compared to a single-view approach, as evidenced by Fig. 6.4. This claim is also supported by the reported performance of the DoG-Detector, which operates independently on the foreground segmented depth observations of each sensor.

The advantages of our generative joint probabilistic approach are also supported by the direct comparison with the proposed end-to-end multi-view CNN architecture. While the reported results are noticeable, considering that the neural network is only trained with synthetic images, the overall performance of the proposed CNN is not competitive. This is an interesting result, since both approaches are based on the same generative scene model (except the additional randomization) and do not use any further supervision (e. g. labeled real-world image data). Thus, the CNN approach can be interpreted as black box inference method, replacing the more involved, hand crafted mean-field variational inference optimization. We emphasize that while the former is a discriminative method the latter is a generative method. In view of the reported results, we can conclude that for multi-view people detection in overlapping depth images it is worth to take the extra effort to deduce the mean-field equations. In spite of the reported results, we believe that by leveraging a large scale training data set, it might be possible to achieve competitive results with a comparable CNN architecture. However, a suitable large-scale data set covering multi-view people detection in top-view depth images does not exist yet.

Even though the mean-field approximation applied to a single frame already achieves quite strong performance, the results reported in Sect. 6.4.2 indicate that leveraging the temporal context can further increase the detection performance. On the test sequence the best performance was achieved by the temporal smoothing approach, directly followed by the Bayesian filtering approach. Considering the rather moderate quantitative performance improvements on an evaluation sequence of limited length, the results have to be interpreted carefully. Taking also the qualitative results into account, we can observe several sub-sequences where, compared to the single frame method, the additional temporal information clearly improves the prediction quality. Our hypothesis is that we could observe these effects in a higher frequency on a more challenging data set, including a more cluttered foreground segmentation.

From a practical point of view, the proposed mean-field Bayesian filtering method might be the most promising approach. The temporal smoothing approach achieves the best performance by jointly optimizing over a sequence of temporal views. This has many theoretically appealing aspects and might also be sufficient for some applications. However, for many real-world applications batch processing[11] is often impractical. Therefore, we favor the approximation of the Bayesian filtering distribution for all applications with real-time requirements. Due to the recursive

[11] In contrast to frame-by-frame processing, batch processing refers to a detection method taking a sequence of multi-view frames as input.

definition of the posterior distribution, the method fulfills the requirements of a frame-by-frame detector while almost reaching the detection performance of the temporal smoothing approach. Nonetheless, due to a frame processing time in the order of one second, we have to admit that the current implementation of our approach is not yet real-time capable.

Additionally, we report results for the MAP inference method, based on a continuous latent space. Compared to the exhaustive evaluation of the different manifestation of mean-field approximations, the explanatory power of the MAP inference evaluation is limited. The qualitative results show that for a sufficient initial scene configuration the continuous optimization eventually converges, resulting in a good approximation of the MAP state. However, the continuous optimization highly depends on an adequate initialization. Hence, we see the potential for the MAP inference as a post processing step, to fine-tune rather coarse detections. In practice this can be useful for offline applications with high-precision requirements, e. g. automatic annotation of novel data sets, where initial coarse detections could be given by a human annotator or a detection algorithm such as the proposed mean-field variational inference approach. Since this is a purely offline application, with no real-time requirements, it is also a particular example where the MF-VI temporal smoothing approach would be the favorable choice over the Bayesian filtering method.

7 Conclusion and Future Work

7.1 Conclusion

In the present thesis we have addressed the problem of probabilistic multi-view people detection in overlapping depth images. In particular, we have investigated methods making joint use of the temporal multi-view image evidence. Therefore, we have presented a comprehensive generative probabilistic framework, recasting the problem of multi-view people detection as an inverse problem. The core of this model is a generative scene model, which maps a scene configuration to a synthetic depth observation. The generative approach effectively handles the different appearances of people, leading to a view-point agnostic detector. Since the generative scene model is a function of the projection matrix of each sensor, our framework makes it easy to incorporate a new sensor modality. The only requirement is the definition of an adequate sensor forward model. Based on the proposed generative probabilistic model, we have presented several inference strategies.

For continuous latent space, we have deduced the MAP objective and solved the resulting non-linear least squares optimization problem by leveraging approximate differentiable rendering. While this approach hinges on an adequate initialization, our experiments demonstrate that the MAP inference can be used complementary to the presented methods utilizing a discrete latent space, serving as a fine-tuning post-processing step.

Instead of just estimating a MAP point estimate, we have further investigated mean-field variational inference methods to jointly exploit the multi-view information in order to approximate the probability distribution of people present in the scene. We have deduced the mean-field update equations for the data term and proposed a novel strategy to efficiently approximate the final mean-field expectations by leveraging a pre-computed visual dictionary. For evaluation, we have introduced the novel data set MULTIPLE. The data set is publicly available and in particular covers indoor people detection in overlapping depth images from the top-view. Our experiments have shown state-of-the-art results on the MULTIPLE data set. Even without using the temporal context, we have demonstrated 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.

Furthermore, we have introduced a novel temporal extension of the presented mean-field approach, which leverages the temporal context to regularize the stochastic mean-field optimization process. In particular, we have proposed a grid based dynamics model to describe the flow of probability mass over time, enabling the definition of the joint distribution of people present in the scene across space and time. Based on the joint distribution, we have deduced the mean-field equations for the full joint distribution (temporal smoothing) as well as the recursively defined Bayesian filtering distribution. Our results show that for the temporal smoothing as well as the Bayesian filtering approach, the introduced temporal regularization leads to a more robust estimation of the desired probability distributions, and in consequence increases the detection performance.

For a direct comparison with the proposed probabilistic inference methods, we have additionally introduced an end-to-end multi-view CNN framework. In contrast to prevalent methods in the literature, the CNN architecture is trained only on synthetic depth images, sampled from the (randomized) generative scene model. Although we have reported noticeable results for the proposed CNN framework, the overall performance is not competitive with the proposed mean-field approximations. These results support the claim that, for multi-view people detection in depth images without a large-scale labeled data set, it is worth making use of more involved joint generative probabilistic modeling methods.

In summary, we have introduced a probabilistic framework for indoor people detection in top-view depth images. The overriding goal has been to leverage the temporal multi-view image evidence from all depth sensors jointly to resolve occlusion as well as measurement noise. We have demonstrated that the proposed mean-field methods effectively approximate the joint probability distribution of people present in the scene, leading to state-of-the-art detection performance.

7.2 Future Work

The methods presented in this thesis open up a variety of future research directions. While in this work, we focused on a homogeneous network of passive stereo depth sensors, it would be appealing to integrate different types of sensor modalities into the sensor network. It would be straightforward to integrate other types of depth sensing devices, such as active stereo sensors or time-of-flight cameras. However, as long as an appropriate forward model for the sensor modality can be defined, any kind of sensor can be integrated into the presented frame work. Considering indoor people detection, this could, for example, include monocular video cameras, thermographic cameras or even light barriers.

Besides the integration of new sensor modalities, the proposed forward model for depth sensor could be extended in several ways. A distance and viewpoint dependent noise model could prevent false negative detections, occasionally occurring on the image borders due to heavy measurement noise. Second, a more expressive generative scene model could be employed. For the sake of a low dimensional latent space, in this thesis, a simple rotationally symmetric 3D person model is used. However, one could also extend the latent space and model each individual with a morphable 3D person model, for example incorporating degrees of freedom for height, body size, body shape, and rotation.

From a methodological perspective, there are several opportunities to advance the proposed mean-field optimization. For the approximation of the data mean-field update expectation, we assume that only the direct neighborhood of a grid location affects the expectation. While this assumption is valid for the top-view, for a frontal viewpoint, a more sophisticated approximation would be preferable. This could, for example, be obtained by ray-tracing techniques or by checking for the intersection of the bounding boxes with each other. However, the complexity of the approximated mean-field update expectation increases exponentially with the number of neighbor cells considered. Focusing on the run-time issues in real-time applications, an open question to research is the (partial) parallelization of mean-field updates. The challenge here is to employ parallel mean-field updates, while exhibiting convergence properties similar to the proposed asynchronous mean-field optimization.

While our experiments showed that the overall detection performance can be improved by using the temporal context, there is a great potential to extend this further. For example, the proposed dynamics model considers the distribution of individuals to be conditionally independent, once the previous scene configuration state is given. This assumption makes the problem computationally tractable, but it also significantly limits the expressiveness of the dynamics model. Therefore, a future research direction could be the investigation of different factorizations of the transition distribution. In particular, it would be tempting to employ transition distributions with a pairwise factorization, promising a significant increase in expressiveness while still being tractable. Moreover, it would be engaging to consider more powerful motion models, incorporating the direction and speed of individuals.

Overall, to assess the extensions proposed above, a more extensive quantitative evaluation would be desirable. Therefore, a large scale data set for multi-view people detection in overlapping depth images is needed.

In future work, the MULTIPLE data set could be extended by various real-world scenarios, including customer behavior analysis in retail stores or people tracking at an airport.

A promising, fundamental future research direction will be the methodical combination of generative probabilistic modeling with state-of-the-art data-driven CNN architectures. While this idea has already been proposed in the literature by [9, 10] in the context of multi-view people detection with monocular video sensors, we believe that it is a forward-looking idea and in general will become increasingly interesting for the computer vision community in the upcoming years.

Appendix

A Derivation of Mean-Field Update Equations

A.1 General Mean-Field Equations

The following derivation is based on [70, p. 736 ff.] and [14, p. 465 ff.]. To keep the notation simple, we use $\int f(\mathcal{X})\,d\mathcal{X}$ as short hand for $\int_{\mathcal{D}} \cdots \int_{\mathcal{D}} f(\mathcal{X})\,dx'_1, \cdots\, dx'_m$ and $\int f(\mathcal{X})\,d\mathcal{X}\backslash x'_i$ as the integral over all variables $\{x'_i : i \neq j\}$ respectively. The idea is to derive the update equation for a single distribution q_i by isolating the dependency in the ELBO $\mathcal{L}(q)$.

Assuming the elements $\{q_j : i \neq j\}$ stay fixed, the ELBO for one single distribution can be derived as

$$\mathcal{L}(q_i) = \langle \log p(\mathcal{X}, \mathcal{O}) - \log q(\mathcal{X}) \rangle_{q(\mathcal{X})} \tag{A.1a}$$

$$= \int \prod_j q_j(x'_j) \left(\log p(\mathcal{X}, \mathcal{O}) - \sum_k \log q_k(x'_k) \right) d\mathcal{X} \tag{A.1b}$$

$$= \int \int q_i(x'_i) \prod_{j \neq i} q_j(x'_j) \left(\log p(\mathcal{X}, \mathcal{O}) - \sum_k \log q_k(x'_k) \right) d\mathcal{X} \backslash x'_i \, dx'_i \tag{A.1c}$$

$$= \int q_i(x'_i) \underbrace{\int \prod_{j \neq i} q_j(x'_j) \log p(\mathcal{X}, \mathcal{O}) \, d\mathcal{X} \backslash x'_i \, dx'_i}_{\log \tilde{p}_i(\mathcal{X}, \mathcal{O}) = \langle \log p(\mathcal{X}, \mathcal{O}) \rangle_{q(\mathcal{X} \backslash x'_i)}} \tag{A.1d}$$

$$- \int q_i(x'_i) \int \prod_{j \neq i} q_j(x'_j) \left(\sum_{k \neq i} \log q_k(x'_k) + \log q_i(x'_i) \right) d\mathcal{X} \backslash x'_i \, dx'_i$$

$$= \int q_i(x'_i) \log \tilde{p}_i(\mathcal{X}, \mathcal{O}) \, dx'_i - \int q_i(x'_i) \log q_i(x'_i) \, dx'_i + \text{const} \tag{A.1e}$$

$$= \int \log \frac{\tilde{p}_i(\mathcal{X}, \mathcal{O})}{q_i(x'_i)} q_i(x'_i) \, dx'_i + \text{const}. \tag{A.1f}$$

Up to a constant factor the ELBO $\mathcal{L}(q_i)$ is equivalent to the negative KL divergence, thus we can re-write the objective as

$$\mathcal{L}(q_i) = -\mathrm{KL}(q_i \parallel \tilde{p}_i), \tag{A.2}$$

with

$$\tilde{p}_i(\mathcal{X}, \mathcal{O}) = \frac{1}{Z_i} \exp\Big(\langle \log p(\mathcal{X}, \mathcal{O}) \rangle_{q(\mathcal{X} \backslash x'_i)} \Big). \tag{A.3}$$

It follows that the optimal solution for the objective $\hat{q}_i(x_i') = \arg\max \mathcal{L}(q_i)$ is given as

$$q_i(x_i') = \frac{1}{Z_i} \exp\Big(\langle \log p(\mathcal{X}, \mathcal{O}) \rangle_{q(\mathcal{X} \backslash x_i')} \Big). \tag{A.4}$$

A.2 POM Mean-Field Equations

Inserting the probabilistic model of POM to the mean-field expectation we obtain the unconditioned expectation

$$\begin{aligned} \langle \log p(\mathbf{b}, \mathbf{x}) \rangle_{q(\mathbf{x} \backslash x_i)} &= \langle \log p(\mathbf{b} \mid \mathbf{x}) p(\mathbf{x}) \rangle_{q(\mathbf{x} \backslash x_i)} && \text{(A.5)} \\ &= -\sum_{c=1}^{C} \big\langle \delta_{\text{pom}}(b_c, A_c(\mathbf{x})) \big\rangle_{q(\mathbf{x} \backslash x_i)} && \text{(A.6)} \\ &\quad + \langle \log p(\mathbf{x}) \rangle_{q(\mathbf{x} \backslash x_i)} \, . \end{aligned}$$

By inserting (A.6) in (3.36), the POM mean-field equations are given as

$$
\begin{aligned}
\hat{q}_i(x_i = 1) &= \frac{\exp\Big(\langle \log p(\mathbf{b}, \mathbf{x} \mid x_i = 1)\rangle_{q(\mathbf{x}\setminus x_i)}\Big)}{\sum_{s\in\{0,1\}} \exp\Big(\langle \log p(\mathbf{b}, \mathbf{x} \mid x_i = s)\rangle_{q(\mathbf{x}\setminus x_i)}\Big)} && \text{(A.7a)}\\
&= \Bigg[1 + \exp\Bigg(\langle \log p(\mathbf{b}, \mathbf{x} \mid x_i = 0)\rangle_{q(\mathbf{x}\setminus x_i)} && \text{(A.7b)}\\
&\qquad - \langle \log p(\mathbf{b}, \mathbf{x} \mid x_i = 1)\rangle_{q(\mathbf{x}\setminus x_i)}\Bigg)\Bigg]^{-1}\\
&= \Bigg[1 + \exp\Bigg(\langle \log p(\mathbf{x} \mid x_i = 0)\rangle_{q(\mathbf{x}\setminus x_i)} && \text{(A.7c)}\\
&\qquad - \sum_{c=1}^{C} \big\langle \delta_{\text{pom}}(b_c, A_c(\mathbf{x}|x_i = 0))\big\rangle_{q(\mathbf{x}\setminus x_i)}\\
&\qquad - \langle \log p(\mathbf{x} \mid x_i = 1)\rangle_{q(\mathbf{x}\setminus x_i)}\\
&\qquad + \sum_{c=1}^{C} \big\langle \delta_{\text{pom}}(b_c, A_c(\mathbf{x}|x_i = 1))\big\rangle_{q(\mathbf{x}\setminus x_i)}\Bigg)\Bigg]^{-1}\\
&= \Bigg[1 + \exp\Bigg(\bigg\langle \log \frac{p(\mathbf{x} \mid x_i = 0)}{p(\mathbf{x} \mid x_i = 1)}\bigg\rangle_{q(\mathbf{x}\setminus x_i)} && \text{(A.7d)}\\
&\qquad + \sum_{c=1}^{C} \Big(\big\langle \delta_{\text{pom}}(b_c, A_c(\mathbf{x}|x_i = 1))\big\rangle_{q(\mathbf{x}\setminus x_i)}\\
&\qquad\qquad - \big\langle \delta_{\text{pom}}(b_c, A_c(\mathbf{x}|x_i = 0))\big\rangle_{q(\mathbf{x}\setminus x_i)}\Big)\Bigg)\Bigg]^{-1}.
\end{aligned}
$$

B Derivation of the General Probabilistic Model

In this appendix we provide the derivations for Sect. 4.2. The derivations are based on the joint model defined in (4.7), restated as

$$p(\mathcal{X}_{1:T}, \mathbf{o}_{1:T}) = p(\mathbf{o}_1 \mid \mathcal{X}_1) p(\mathcal{X}_1) \prod_{t=2}^{T} p(\mathbf{o}_t \mid \mathcal{X}_t) p(\mathcal{X}_t \mid \mathcal{X}_{t-1}) \,. \quad \text{(B.1)}$$

According to Sect. 4.2, in the following we omit the integration domain, where $\int p(\mathcal{X}_{1:T}) \, d\mathcal{X}_{1:T}$ is a short hand for $\int_{\mathcal{D}} \cdots \int_{\mathcal{D}} p(\mathcal{X}_{1:T}) \, d\mathcal{X}_1 \ldots d\mathcal{X}_T$ with $\mathcal{D}$ defined as the full domain of $\mathcal{X}$. The integrals are defined over an abstract state space implied by $\mathcal{X}$. Depending on the concrete manifestation of the scene configuration space, the integrals have to be refined or, in case of a discrete scene configuration space, replaced by the sum over all discrete states.

B.1 Posterior Distribution

For $t \in \{2, \ldots, T\}$:

$$p(\mathcal{X}_{1:t} \mid \mathbf{o}_{1:t}) = \frac{p(\mathbf{o}_{1:t}, \mathcal{X}_{1:t})}{p(\mathbf{o}_{1:t})} \tag{B.2a}$$

$$= \frac{\prod_{k=2}^{t} p(\mathbf{o}_k \mid \mathcal{X}_k) p(\mathcal{X}_k \mid \mathcal{X}_{k-1})}{p(\mathbf{o}_{1:t})} \tag{B.2b}$$

$$= \frac{p(\mathbf{o}_t \mid \mathcal{X}_t) p(\mathcal{X}_t \mid \mathcal{X}_{t-1}) p(\mathbf{o}_{1:t-1}, \mathcal{X}_{1:t-1})}{p(\mathbf{o}_t \mid \mathbf{o}_{1:t-1}) p(\mathbf{o}_{1:t-1})} \tag{B.2c}$$

$$= \frac{p(\mathbf{o}_t \mid \mathcal{X}_t) p(\mathcal{X}_t \mid \mathcal{X}_{t-1}) p(\mathcal{X}_{1:t-1} \mid \mathbf{o}_{1:t-1})}{p(\mathbf{o}_t \mid \mathbf{o}_{1:t-1})} \,. \tag{B.2d}$$

B.2 Predicted Likelihood

For $t \in \{2, \ldots, T\}$:

$$p(\mathbf{o}_t \mid \mathbf{o}_{1:t-1}) = \int p(\mathbf{o}_t \mid \mathcal{X}_t) p(\mathcal{X}_t \mid \mathcal{X}_{t-1}) p(\mathcal{X}_{1:t-1} \mid \mathbf{o}_{1:t-1}) \, d\mathcal{X}_{1:t} \tag{B.3a}$$

$$= \int p(\mathbf{o}_t \mid \mathcal{X}_t) \underbrace{\int p(\mathcal{X}_t \mid \mathcal{X}_{t-1}) p(\mathcal{X}_{t-1} \mid \mathbf{o}_{1:t-1}) \, d\mathcal{X}_{t-1}}_{p(\mathcal{X}_t \mid \mathbf{o}_{1:t-1})} \, d\mathcal{X}_t \tag{B.3b}$$

$$= \int p(\mathbf{o}_t \mid \mathcal{X}_t) p(\mathcal{X}_t \mid \mathbf{o}_{1:t-1}) d\mathcal{X}_t \,. \tag{B.3c}$$

B.3 Filtering Distribution

For $t \in \{2, \dots, T\}$:

$$\begin{aligned}
p(\mathcal{X}_t \mid \mathbf{o}_{1:t}) &= \int p(\mathcal{X}_{1:t} \mid \mathbf{o}_{1:t}) \, d\mathcal{X}_{1:t-1} && \text{(B.4a)} \\
&= \underbrace{\frac{p(\mathbf{o}_t \mid \mathcal{X}_t)}{p(\mathbf{o}_t \mid \mathbf{o}_{1:t-1})}}_{\nu} \int p(\mathcal{X}_t \mid \mathcal{X}_{t-1}) p(\mathcal{X}_{1:t-1} \mid \mathbf{o}_{1:t-1}) \, d\mathcal{X}_{1:t-1} && \text{(B.4b)} \\
&= \nu \int p(\mathcal{X}_t \mid \mathcal{X}_{t-1}) \underbrace{\int p(\mathcal{X}_{1:t-1} \mid \mathbf{o}_{1:t-1}) \, d\mathcal{X}_{1:t-2}}_{p(\mathcal{X}_{t-1} | \mathbf{o}_{1:t-1})} \, d\mathcal{X}_{t-1} && \text{(B.4c)} \\
&= \nu \underbrace{\int p(\mathcal{X}_t \mid \mathcal{X}_{t-1}) p(\mathcal{X}_{t-1} \mid \mathbf{o}_{1:t-1}) \, d\mathcal{X}_{t-1}}_{p(\mathcal{X}_t | \mathbf{o}_{1:t-1})} . && \text{(B.4d)}
\end{aligned}$$

Bibliography

[1] **Ahmad, M. M., Ahmed, I., Ullah, K., and Ahmad, M. M.** *A deep neural network approach for top view people detection and counting*. In: *Proceedings of the IEEE 10th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*. 2019.

[2] **Ahmad, M., Ahmed, I., and Adnan, A.** *Overhead view person detection using YOLO*. In: *Proceedings of the IEEE 10th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*. 2019.

[3] **Ahmad, M., Ahmed, I., Ullah, K., Khan, I., Khattak, A., and Adnan, A.** *Person detection from overhead view: A survey*. In: *International Journal of Advanced Computer Science and Applications (IJACSA)* 10.4 (2019).

[4] **Ahmed, I. and Adnan, A.** *A robust algorithm for detecting people in overhead views*. In: *Cluster Computing* 21.1 (2017), pp. 633–654.

[5] **Ahmed, I., Ahmad, M., Ahmad, A., and Jeon, G.** *Top view multiple people tracking by detection using deep SORT and YOLOv3 with transfer learning: within 5G infrastructure*. In: *International Journal of Machine Learning and Cybernetics* (2020).

[6] **Alahi, A., Jacques, L., Boursier, Y., and Vandergheynst, P.** *Sparsity driven people localization with a heterogeneous network of cameras*. In: *Journal of Mathematical Imaging and Vision* 41.1-2 (2011), pp. 39–58.

[7] **Anjum, N. and Cavallaro, A.** *Trajectory association and fusion across partially overlapping cameras*. In: *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 2009.

[8] **Bagautdinov, T., Fleuret, F., and Fua, P.** *Probability occupancy maps for occluded depth images*. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.

[9] **Baqué, P. B.** *Mean-Field methods for structured deep-learning in computer vision*. PhD thesis. École polytechnique fédérale de Lausanne (EPFL), 2018.

[10] **Baqué, P., Fleuret, F., and Fua, P.** *Deep occlusion reasoning for multi-camera multi-target detection*. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017.

[11] **Barber, D.** *Bayesian reasoning and machine learning*. Cambridge University Press, 2011.

[12] **Berclaz, J., Shahrokni, A., Fleuret, F., Ferryman, J., and Fua, P.** *Evaluation of probabilistic occupancy map people detection for surveillance systems*. In: *Proceedings of the IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*. 2009.

[13] **Beyer, L., Breuers, S., Kurin, V., and Leibe, B.** *Towards a principled integration of multi-camera re-identification and tracking through optimal Bayes filters*. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[14] **Bishop, C. M.** *Pattern recognition and machine learning*. Springer, 2006.

[15] **Blei, D. M., Kucukelbir, A., and McAuliffe, J. D.** *Variational inference: A review for statisticians*. In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877.

[16] **Bohacek, E., Coates, A. J., and Selviah, D. R.** *Volumetric calculation of quantization error in 3-D vision systems*. 2020. arXiv: 2010.08390 `[cs.CV]`.

[17] **Camplani, M., Paiement, A., Mirmehdi, M., Damen, D., Hannuna, S., Burghardt, T., and Tao, L.** *Multiple human tracking in RGB-depth data: A survey*. In: *IET Computer Vision* 11.4 (2017), pp. 265–285.

[18] **Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y.** *OpenPose: Realtime multi-person 2D pose estimation using part affinity fields*. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 43.1 (2021), pp. 172–186.

[19] **Carletti, V., Del Pizzo, L., Percannella, G., and Vento, M.** *An efficient and effective method for people detection from top-view depth cameras*. In: *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 2017.

[20] **Carraro, M., Munaro, M., Burke, J., and Menegatti, E.** *Real-time marker-less multi-person 3D pose estimation in RGB-depth camera networks*. In: *Proceedings of the International Conference on Intelligent Autonomous Systems*. 2019.

[21] **Castellano, G., Castiello, C., Cianciotta, M., Mencar, C., and Vessio, G.** *Multi-view convolutional network for crowd counting in drone-captured images*. In: *Proceedings of the Workshop of the European Conference on Computer Vision (ECCV)*. 2020.

[23] **Chavdarova, T. and Fleuret, F.** *Deep multi-camera people detection*. In: *Proceedings of the IEEE International Conference on Machine Learning and Applications (ICMLA)*. 2017.

[22] **Chavdarova, T., Baqué, P., Bouquet, S., Maksai, A., Jose, C., Bagautdinov, T., Lettry, L., Fua, P., Van Gool, L., and Fleuret, F.** *WILDTRACK: A multi-camera HD dataset for dense unscripted pedestrian detection*. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[24] **Chen, M., Chen, Y., Loc, T. T., and Ni, B.** *Real-time multiple pedestrians tracking in multi-camera system*. In: *Proceedings of the International Conference on Multimedia Modeling*. 2020.

[25] **Cipra, B. A.** *An introduction to the Ising model*. In: *The American Mathematical Monthly* 94.10 (1987), p. 937.

[26] **Cover, T. M. and Thomas, J. A.** *Elements of information theory*. 2. ed. Wiley-Interscience, 2006.

[27] **Dalal, N. and Triggs, B.** *Histograms of oriented gradients for human detection*. In: *Proceedings of the IEEE International Conference on Pattern Recognition (CVPR)*. 2005.

[28] **David M. Blei and Andrew Y. Ng and Michael I. Jordan and John Lafferty**. *Latent Dirichlet allocation*. In: *Journal of Machine Learning Research* 3 (2003), pp. 993–1022.

[29] **Davis, J. and Goadrich, M.** *The relationship between precision-recall and ROC curves*. In: *Proceedings of the 23rd International Conference on Machine Learning (ICML)*. 2006.

[30] **Del Pizzo, L., Foggia, P., Greco, A., Percannella, G., and Vento, M.** *Counting people by RGB or depth overhead cameras*. In: *Pattern Recognition Letters* 81 (2016), pp. 41–50.

[31] **Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, and Li Fei-Fei**. *ImageNet: A large-scale hierarchical image database*. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2009.

[32] **Dollar, P., Wojek, C., Schiele, B., and Perona, P.** *Pedestrian detection: A benchmark*. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2010.

[33] **Ertler, C., Possegger, H., Opitz, M., and Bischof, H.** *Pedestrian detection in RGB-D images from an elevated viewpoint*. In: *Proceedings of the 22nd Computer Vision Winter Workshop*. 2017.

[34] **Eshel, R. and Moses, Y.** *Homography based multiple camera detection and tracking of people in a dense crowd*. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2008.

[35] **Fernando, T., Denman, S., Sridharan, S., and Fookes, C.** *Tracking by prediction: A deep generative model for mutli-person localisation and tracking*. In: *Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2018.

[36] **Fischer, Y. and Beyerer, J.** *A top-down-view on intelligent surveillance systems*. In: *Proceedings of the International Conference on Systems*. 2012.

[37] **Fleet, D. J.** *Motion models for people tracking*. In: *Visual Analysis of Humans: Looking at People*. Springer, 2011.

[39] **Fleuret, F., Lengagne, R., and Fua, P.** *Fixed point probability field for complex occlusion handling*. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2005.

[38] **Fleuret, F., Berclaz, J., Lengagne, R., and Fua, P.** *Multicamera people tracking with a probabilistic occupancy map*. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 30.2 (2008), pp. 267–282.

[40] **Fox, C. W. and Roberts, S. J.** *A tutorial on variational Bayesian inference*. In: *Artificial Intelligence Review* 38.2 (2012), pp. 85–95.

[41] **Fuentes-Jimenez, D., Martin-Lopez, R., Losada-Gutierrez, C., Casillas-Perez, D., Macias-Guarasa, J., Luna, C. A., and Pizarro, D.** *DPDnet: A robust people detector using deep learning with an overhead depth camera*. In: *Expert Systems with Applications* 146 (2020), p. 113168.

[42] **Ge, W. and Collins, R. T.** *Crowd detection with a multiview sampler*. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2010.

[43] **Green, P. J.** *Reversible jump Markov chain Monte Carlo computation and Bayesian model determination*. In: *Biometrika* 82.4 (1995), pp. 711–732.

[44] **Griewank, A. and Walther, A.** *Evaluating derivatives: Principles and techniques of algorithmic differentiation*. 2nd edt. Society for Industrial and Applied Mathematics, 2008.

[45] **Hacinecipoglu, A., Konukseven, E. I., and Koku, A. B.** *Pose invariant people detection in point clouds for mobile robots*. In: *International Journal of Mechanical Engineering and Robotics Research* 9.5 (2020), pp. 709–715.

[46] **Hou, L., Wan, W., Hwang, J. N., Muhammad, R., Yang, M., and Han, K.** *Human tracking over camera networks: A review*. In: *Eurasip Journal on Advances in Signal Processing* 2017.1 (2017), p. 43.

[47] **Hou, Y., Zheng, L., and Gould, S.** *Multiview detection with feature perspective transformation*. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2020.

[48] **Iguernaissi, R., Merad, D., Aziz, K., and Drap, P.** *People tracking in multi-camera systems: A review*. In: *Multimedia Tools and Applications* 78.8 (2019), pp. 10773–10793.

[49] **Ising, E.** *Beitrag zur Theorie des Ferromagnetismus*. In: *Zeitschrift für Physik* 31.1 (1925), pp. 253–258.

[50] **Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K.** *Introduction to variational methods for graphical models*. In: *Machine Learning* 37.2 (1999), pp. 183–233.

[51] **Kamberova, G. and Bajcsy, R.** *Sensor errors and uncertainties in stereo reconstruction*. In: *Empirical Evaluation Techniques for Computer Vision* (1998), pp. 96–116.

[52] **Kayumbi, G., Anjum, N., and Cavallaro, A.** *Global trajectory reconstruction from distributed visual sensors*. In: *Proceedings of the 2nd ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)* (2008).

[53] **Khan, S. M. and Shah, M.** *Tracking multiple occluding people by localizing on multiple scene planes*. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 31.3 (2009), pp. 505–519.

[54] **Kieritz, H., Becker, S., Hubner, W., and Arens, M.** *Online multi-person tracking using integral channel features*. In: *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 2016.

[55] **Kim, W. and Jung, C.** *Illumination-invariant background subtraction: Comparative review, models, and prospects*. In: *IEEE Access* 5 (2017), pp. 8369–8384.

[56] **Kingma, D. P. and Ba, J.** *Adam: A method for stochastic optimization*. In: *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. 2015.

[57] **Klenke, A.** *Wahrscheinlichkeitstheorie*. 4th edt. Springer, 2020.

[58] **Korkalo, O., Tikkanen, T., Kemppi, P., and Honkamaa, P.** *Auto-calibration of depth camera networks for people tracking*. In: *Machine Vision and Applications* 30.4 (2019), pp. 671–688.

[59] **Kulkarni, T. D., Kohli, P., Tenenbaum, J. B., and Mansinghka, V.** *Picture: A probabilistic programming language for scene perception*. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.

[60] **Kullback, S. and Leibler, R. A.** *On information and sufficiency*. In: *Annals of Mathematical Statistics* 22 (1951), pp. 79–86.

[61] **Li, H., Liu, J., Zhang, G., Gao, Y., and Wu, Y.** *Multi-glimpse LSTM with color-depth feature fusion for human detection*. In: *Proceedings of the IEEE international Conference on Image Processing (ICIP)*. 2018.

[62] **Li, J., Liang, X., Shen, S. M., Xu, T., Feng, J., and Yan, S.** *Scale-aware fast R-CNN for pedestrian detection*. In: *IEEE Transactions on Multimedia* 20.4 (2017), pp. 1–10.

[63] **Liciotti, D., Paolanti, M., Frontoni, E., and Zingaretti, P.** *People detection and tracking from an RGB-D camera in top-view configuration: Review of challenges and applications*. In: *Proceedings of the International Conference on Image Analysis and Processing (ICIAP)*. 2017.

[64] **Liu, A.-S., Hsu, T.-W., Hsiao, P.-H., Liu, Y.-C., Fu, L.-C., and Fellow, I.** *The manhunt network : People tracking in hybrid- overlapping under the vertical top-view depth camera networks*. In: *Proceedings of the IEEE International Conference on Advanced Robotics and Intelligent Systems (ARIS)*. 2016.

[65] **Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., and Berg, A. C.** *SSD: Single shot multibox detector*. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2016.

[66] **Liu, W., Liao, S., Ren, W., Hu, W., and Yu, Y.** *High-level semantic feature detection: A new perspective for pedestrian detection*. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[67] **Liu, X., Mei, L., Yang, D., Lai, J., and Xie, X.** *Feature visualization based stacked convolutional neural network for human body detection in a depth image*. In: *Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. 2018.

[68] **Loper, M. M. and Black, M. J.** *OpenDR: An approximate differentiable renderer*. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2014.

[69] **Munaro, M., Basso, F., and Menegatti, E.** *OpenPTrack: Open source multi-camera calibration and people tracking for RGB-D camera networks*. In: *Robotics and Autonomous Systems* 75 (2016), pp. 525–538.

[70] **Murphy, K. P.** *Machine learning: A probabilistic perspective*. MIT Press, 2012.

[71] **Nocedal, Jorge and Wright, S. J.** *Numerical optimization*. 2nd edt. Springer, 2006.

[72] **Otero, J. and Sánchez, L.** *Soft methods for bounding the uncertainty of stereo calibration and triangulation*. In: *Proceedings of the IEEE International Conference on Fuzzy Systems*. 2013.

[73] **Pang, Y., Xie, J., Khan, M. H., Anwer, R. M., Khan, F. S., and Shao, L.** *Mask-guided attention network for occluded pedestrian detection*. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2019.

[74] **Parisi, G.** *Statistical field theory*. Addison-Wesley, 1988.

[75] **Peng, P., Tian, Y., Wang, Y., Li, J., and Huang, T.** *Robust multiple cameras pedestrian detection with multi-view Bayesian network*. In: *Pattern Recognition* 48.5 (2015), pp. 1760–1772.

[76] **Powell, M. J. D.** *A hybrid method for nonlinear equations*. In: *Numerical Methods for Nonlinear Algebraic Equations*. Gordon and Breach, 1970.

[77] **Powell, M. J. D.** *A new algorithm for unconstrained optimization*. In: *Nonlinear Programming*. Academic Press, 1970.

[78] **Qi, C. R., Liu, W., Wu, C., Su, H., and Guibas, L. J.** *Frustum PointNets for 3D object detection from RGB-D data*. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[79] **Rahimi, A., Dunagan, B., and Darrell, T.** *Simultaneous calibration and tracking with a network of non-overlapping sensors*. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2004.

[80] **Rauter, M.** *Reliable human detection and tracking in top-view depth images*. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2013.

[81] **Redmon, J. and Farhadi, A.** *YOLO9000: Better, faster, stronger*. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[82] **Redmon, J. and Farhadi, A.** *YOLOv3: An incremental improvement*. 2018. arXiv: 1804.02767 `[cs.CV]`.

[83] **Ren, S., He, K., Girshick, R., and Sun, J.** *Faster R-CNN: Towards real-time object detection with region proposal networks*. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 39.6 (2017), pp. 1137–1149.

[84] **Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., and Savarese, S.** *Generalized intersection over union: A metric and a loss for bounding box regression*. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[85] **Ristani, E. and Tomasi, C.** *Features for multi-target multi-camera tracking and re-identification*. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[86] **Rusu, R. B., Bradski, G., Thibaux, R., and Hsu, J.** *Fast 3D recognition and pose using the viewpoint feature histogram*. In: *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*. 2010.

[87] **Saito, T. and Rehmsmeier, M.** *The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets*. In: *PLOS ONE* 10.3 (2015).

[88] **Sankaranarayanan, A. C., Veeraraghavan, A., and Chellappa, R.** *Object detection, tracking and recognition for multiple smart cameras*. In: *Proceedings of the IEEE* 96.10 (2008), pp. 1606–1624.

[89] **Saputra, M. R. U., Widyawan, Putra, G. D., and Santosa, P. I.** *Indoor human tracking application using multiple depth-cameras*. In: *Proceedings of the International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. 2012.

[90] **Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.** *Dropout: A simple way to prevent neural networks from overfitting*. In: *Journal of Machine Learning Research* 15.1 (2014), pp. 1929–1958.

[91] **Sun, S. W., Kuo, C. H., and Chang, P. C.** *People tracking in an environment with multiple depth cameras: A skeleton-based pairwise trajectory matching scheme*. In: *Journal of Visual Communication and Image Representation* 35 (2016), pp. 36–54.

[92] **Sun, S., Akhtar, N., Song, H., Zhang, C., Li, J., and Mian, A.** *Benchmark data and method for real-time people counting in cluttered scenes using depth sensors*. In: *IEEE Transactions on Intelligent Transportation Systems* 20.10 (2019), pp. 3599–3612.

[93] **Tang, Z., Gu, R., and Hwang, J.** *Joint multi-view people tracking and pose estimation for 3D scene reconstruction*. In: *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*. 2018.

[94] **Thrun, S.** *Learning occupancy grid maps with forward sensor models*. In: *Autonomous Robots* 15.2 (2003), pp. 111–127.

[95] **Tian, L., Li, M., Hao, Y., Liu, J., Zhang, G., and Chen, Y. Q.** *Robust 3-D human detection in complex environments with a depth camera*. In: *IEEE Transactions on Multimedia* 20.9 (2018), pp. 2249–2261.

[96] **Tseng, T. E., Liu, A. S., Hsiao, P. H., Huang, C. M., and Fu, L. C.** *Real-time people detection and tracking for indoor surveillance using multiple top-view depth cameras*. In: *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*. 2014.

[97] **Wainwright, M. J. and Jordan, M. I.** *Graphical models, exponential families, and variational inference*. In: *Foundations and Trends in Machine Learning* 1.1-2 (2008), pp. 1–305.

[98] **Wang, X.** *Intelligent multi-camera video surveillance: A review*. In: *Pattern Recognition Letters* 34.1 (2013), pp. 3–19.

[99] **Wei, S. E., Ramakrishna, V., Kanade, T., and Sheikh, Y.** *Convolutional pose machines*. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[100] **Wojke, N., Bewley, A., and Paulus, D.** *Simple online and realtime tracking with a deep association metric*. In: *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. 2017.

[101] **Xu, Y., Liu, X., Liu, Y., and Zhu, S. C.** *Multi-view people tracking via hierarchical trajectory composition*. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[102] **You, Q. and Jiang, H.** *Real-time 3D deep multi-camera tracking*. 2020. arXiv: 2003.11753.

[103] **Zhang, Q. and Chan, A. B.** *Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns*. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[104] **Zhang, S., Benenson, R., Omran, M., Hosang, J., and Schiele, B.** *Towards reaching human performance in pedestrian detection*. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 40.4 (2018), pp. 973–986.

[105] **Zhang, X., Yan, J., Feng, S., Lei, Z., Yi, D., and Li, S. Z.** *Water filling: Unsupervised people counting via vertical kinect sensor*. In: *Proceedings of the IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*. 2012.

[106] **Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., and Torr, P. H. S.** *Conditional random fields as recurrent neural networks*. In: *Proceedings of the International Conference on Computer Vision (ICCV)*. 2015.

[107] **Zhou, F., Wang, H., Yang, Z., and Xie, D.** *A novel 3D head multi-feature constraint method for human localization based on multiple depth cameras*. In: *Proceedings of the International Conference on Intelligent Manufacturing and Internet of Things (IMIOT)*. 2018.

List of Publications

[108] **Wetzel, J., Laubenheimer, A., and Heizmann, M.** *Joint probabilistic people detection in overlapping depth images*. In: *IEEE Access* 8 (2020), pp. 28349–28359.

[109] **Wetzel, J., Laubenheimer, A., and Heizmann, M.** *Temporal smoothing for joint probabilistic people detection in a depth sensor network*. In: *Proceedings of the IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. 2020.

[110] **Wetzel, J., Zeitvogel, S., Laubenheimer, A., and Heizmann, M.** *Towards global people detection and tracking using multiple depth sensors*. In: *Proceedings of the IEEE International Symposium on Electronics and Telecommunications (ISETC)*. 2018.

[111] **Wetzel, J., Zeitvogel, S., Laubenheimer, A., and Heizmann, M.** *Scene-adaptive optimization scheme for depth sensor networks*. In: *Proceedings of the 5th Collaborative European Research Conference (CERC)*. 2019.

[112] **Wetzel, J., Zeitvogel, S., Laubenheimer, A., and Heizmann, M.** *People detection in a depth sensor network via multi-view CNNs trained on synthetic data*. In: *Proceedings of the IEEE International Symposium on Electronics and Telecommunications (ISETC)*. 2020.