Image based multi-level environment analysis
(Képi alapú többszintű környezetelemzés)

A thesis submitted for the
Doctoral Degree of the Hungarian Academy of Sciences
(D.Sc.)

Csaba Benedek, Ph.D.

Institute for Computer Science and Control
Hungarian Academy of Sciences

Budapest, 2019

dc_942_14

Acknowledgements

I gratefully acknowledge the help of my closest colleagues at the Machine Perception Research Laboratory (MPLab) of the Institute for Computer Science and Control of Hungarian Academy of Sciences (MTA SZTAKI), giving always very significant impacts on my ongoing works. I pay special thanks to Professor Tamás Szirányi, the head of the laboratory, my former supervisor and mentor for more than 15 years.

Several B.Sc., M.Sc. and Ph.D. students working with my supervision from the Pázmány Péter Catholic University (PPCU) and the Budapest University of Technology and Economics (BME) have greatly contributed to the conducted research. I thank all of them, especially my Ph.D. students Attila Börcs, Balázs Nagy and Yahya Ibrahim, and among my many talented M.Sc. students Bence Gálai and Oszkár Józsa who contributed as co-authors to several joint publications.

I also thank on both professional and personal levels to my senior and doctoral student colleagues in MPLab: Dániel Baráth, Iván Eichhardt, Levente Hajder, László Havasi, Zsolt Jankó, Anita Keszler, Ákos Kiss, Attila Kiss, Levente Kovács, András Kriston, András Majdik, Andrea Manno-Kovács, Zoltán Pusztai, Zoltán Rózsa, Maha Shadaydeh, László Spórás, Zoltán Szlávik, László Tizedes, Domonkos Varga, Ákos Utasi. Thanks to Mónika Barti and Anikó Vágvölgyi from the administration staff, and to all undergraduates and software developers of the laboratory. I would like to thank Eszter Nagy from the SZTAKI Library for help in organizing my conference trips and for managing our publication records for several years.

My colleagues from various collaborating institutes also provided valuable contributions: Josiane Zerubia and Xavier Descombes from INRIA Sophia Antipolis France, Marco Martorella and Fabrizio Berizzi from the University of Pisa, Zoltán Kató from the University of Szeged, László Jakab and Olivér Krammer from the Department of Electronic Technology of BME, and Dmitry Chetverikov from MTA SZTAKI. I thank Anuj Srivastava for hosting me at the Florida State University as a postdoc visitor.

I also thank the deans of the PPCU Faculty of Information Technology and Bionics, Judit Nyékyné Gaizler, Péter Szolgay and Kristóf Iván, and department heads at BME,

Gábor Harsányi, László Szirmay-Kalos and Bálint Kiss for giving me the opportunities to lecture and supervise students in the universities.

During my research, I had the opportunity to work with a large variety of particular test data. The remotely sensed satellite images and Lidar point clouds were provided by the Airbus Defence and Space Hungary Ltd, the French Defense Ministry and the Liama Sino-French Laboratory. I received radar images from the university of Pisa, Italy. Obtaining our own mobile and terrestrial Lidar laser scanners in the MPLab was possible through the internal R&D grant of MTA SZTAKI, and the Infrastructure Grant of the Hungarian Academy of Sciences, respectively. We received further laser scanning data from Budapest Közút Zrt.

The work introduced in this thesis was partially supported by various projects and grants: Janos Bolyai Research Fellowship of the Hungarian Academy of Sciences, researcher initiated projects of the National Research, Development and Innovation Fund (grants NKFIA K-120233 and KH-125681), the Széchenyi 2020 Program at PPCU (grant EFOP-3.6.2-16-2017-00013), the DUSIREF project of the European Space Agency under the PECS-HU framework, the PROACTIVE EU FP7-SECURITY Project, the i4D (intergated4D) project of MTA SZTAKI, the *Comprehensive Remote Sensing Data Analysis* postdoctoral project of the Hungarian Scientific Research Fund (grant OTKA-101598), APIS and MEDUSA European Defence Agency (EDA) projects, and the New Hungary Development Plan (TÁMOP) project at BME.

I am very grateful to my lovely Lívi and Lóci, to my whole family and to all of my friends who always believed in me and supported me in all possible ways.

Abstract

In this thesis novel approaches are proposed for multi-level scene interpretation based on various 2D and 3D imaging sources. We focus on measurements of up-top-date optical cameras, radars and laser scanners both in terrestrial and airborne configurations. Our central aim is to explore common problems appearing in different application domains, and address them by joint methodological approaches. To ensure the theoretical basis of the new models, the surveys and algorithmic developments are performed in well established Bayesian frameworks, or use recent results of machine learning research. Low level scene understanding functions are formulated as various image segmentation problems, where we take the advantages of the Markov Random Field (MRF) probabilistic framework, which admits us to consider in parallel data-dependent and prior constraints, to get smooth, noiseless, and observation consistent classification outputs. At object level scene analysis, we rely on the literature of Marked Point Process (MPP) approaches, which consider strong geometric and prior interaction constraints in object population modeling. Particularly, we introduce key developments in spatial hierarchical decomposition of the observed scenarios, and in temporal extension of complex MRF and MPP models. Additional contributions are also presented in efficient feature extraction, probabilistic modeling of natural processes and feature integration via local innovations in the model structures. In the last part, we propose new models and algorithms suited to processing the measurements of novel Lidar laser scanners. The research work in this direction enables us to target new application areas, but also implies various new challenges due to the particular measurement characteristics of the sensors. Besides Bayesian techniques, we utilize here the latest deep neural network solutions, fitted to various problems of environment perception. We show by several experiments that the proposed contributions embedded into a strict mathematical toolkit can significantly improve the results in real world 2D/3D test images and videos, for applications on video surveillance, environment monitoring, autonomous driving, remote sensing and optical industrial inspection.

dc_942_14

Contents

List of Figures

dc_942_14

List of Tables

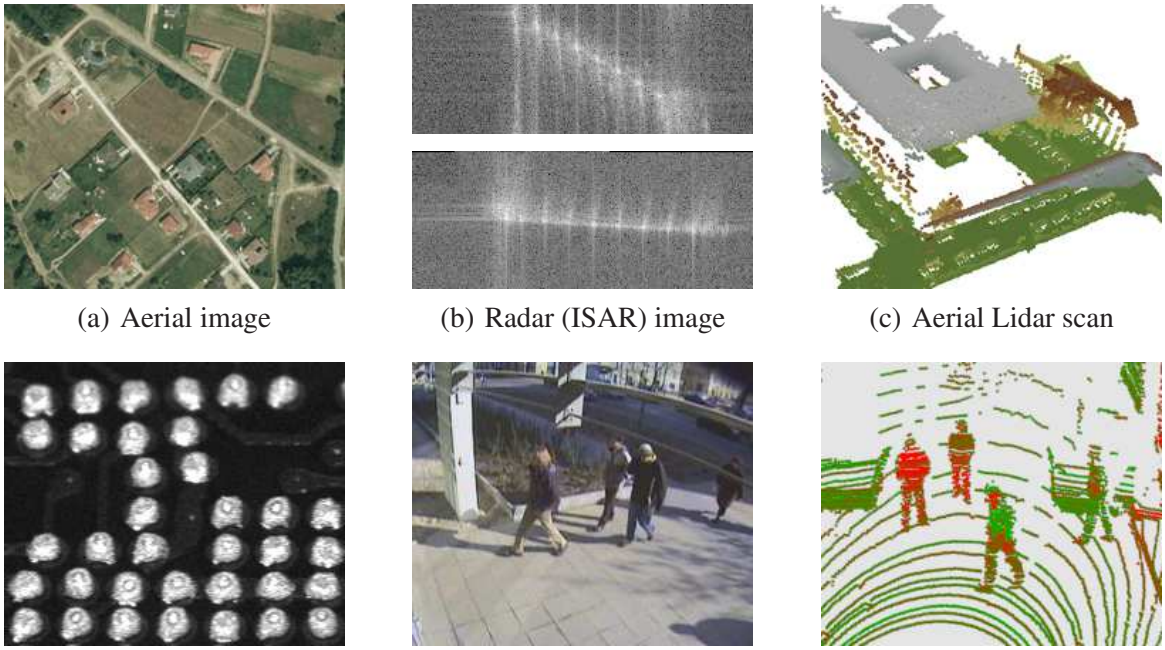
dc_942_14

Chapter 1

Introduction

This thesis deals with selected problems of machine perception, targeting the automated interpretation of the observed static or dynamic environment based on various image-like measurements. Scene understanding is based nowadays far beyond on conventional image processing approaches dealing with standard grayscale or RGB photos. Multi-camera systems, high-speed cameras, radar systems, depth and thermal sensors or laser scanners may be used concurrently to support a given application, therefore proposing a competitive solution for a problem should not only mean to construct the best pattern recognition algorithm but also to chose the best a hardware-software configuration. Some data modalities used in the thesis are demonstrated in Fig. 1.1.

Besides the wide choice of the technologies, we can witness today a quick development of the available sensors in terms of spatial and temporal resolution, number of information channels, level of noise etc. For this reason, by implementing efficient environment perception systems we should answer various challenges of automatic feature extraction, object and event recognition, machine learning, indexing and content-based retrieval. *First*, the developed methodologies should be able to deal with various data sources and they should by highly scalable. This property enables flexible sensor fusion and the replacement of outdated sensors with novel ones providing improved data quality, without complete re-structuring of existing software systems. *Second*, the increased spatial resolution and dimension of the observed data implies that in a single measurement segment one may detect multiple effects on different scales, demanding recognizer algorithms which perform hierarchical interpretation of the content. As an example, in a very high resolution aerial photo, we can jointly analyze macro-level urban or forest regions, separate different districts and roads of the cities, extract and cluster buildings, or focus on smaller objects such as vehicles or street furniture [71, 72]. *Third*, we should also efficiently utilize the multiple available scales of the time dimension. While object motion information can be directly extracted through pixel-by-pixel comparison of the consecutive frames in an image sequence with *video frame rate*; comparing measurements with several months or years time differences captured from the same area needs a high level



(a) Aerial image (b) Radar (ISAR) image (c) Aerial Lidar scan

(d) Printed circuit image (e) Security camera image (f) Terrestrial Lidar scan

Figure 1.1: Examples for different data modalities used in the thesis

modeling approach. The accomplished research work should point therefore towards obtaining a complex system, where the provided information of various data sources is organized into a unified hierarchical scene model, enabling multi-modal representation, recognition and comparison of entities, through combining object level analysis with low level feature extraction.

From a functional point of view, the methods proposed in the thesis present either general preprocessing steps of different early vision applications, or contribute to higher level object based scene analysis modules. In the first case, the introduced models rely on low level local features extracted from the sensor measurements, such as the pixel color values in images, or texture descriptors calculated over small rectangular image parts. The output is a classification (or segmentation) of the observation, which can be interpreted as a semantic labeling of the raw data. For example, in a video frame we can separate the foreground and background pixels, or in an aerial Lidar point cloud roof and terrain regions can be distinguished. Although the classification is primarily based on the extracted local features, which provide posterior (observation dependent) information for the process, additional prior constraints are also exploited to decrease the artifacts due to noise and ambiguities of the input data. One of the simplest, but often used prior condition is the connectivity: we can assume in several problems that the classification should result in homogeneous regions, e.g. in images neighboring pixels correspond usually to the same semantic class.

Markov Random Fields (MRFs) [73] are widely used classification tools since the early eighties, since they are able to simultaneously embed a data model, reflecting the knowledge on the

measurements; and prior constraints, such as spatial smoothness of the solution through a graph based image representation. Since conventional MRFs show some limitations regarding context dependent class modeling, different modified schemes have been recently proposed to increase their flexibility. Triplet Markov fields [74] contain an auxiliary latent process which can be used to describe various subclasses of each class in different manners. Mixed Markov models [75] extend MRFs by admitting data-dependent links between the processing nodes, which fact enables introducing configurable structures in feature integration. Conditional Random Fields (CRFs) directly model the data-driven posterior distributions of the segmentation classes [76]. On their positive points, the above Markovian segmentation approaches are robust and well established for many problems [77]. However, as it will be explained in Chapter 2 in details, the MRF concept offers only a general framework, which has a high degree of freedom. In particularly, a couple of key issues should be efficiently addressed for a given real-world problem. The first one is extracting appropriate features and building a proper probabilistic model of each semantic class. The second key point is developing an appropriate model structure, which consists of simple interactive elements. The arrangement and dialogue of these units is responsible for smoothing the segmentation map or integrating the effects of different features. Choosing the right dimension of the field is also a critical step. MRFs can be either defined on 2D lattices, or on 3D voxel models, but projecting a high dimensional problem to a lower dimensional domain is also a frequently used option. For example, for segmenting a point cloud, a straightforward approach is to construct the MRF in the 3D Euclidean space of the measurements. However, if the point cloud data was recorded by a 2.5D sensor moving on a fixed trajectory, range image representation may provide more efficient results, which is less affected by artifacts of sensor noise and occlusion.

We propose in this thesis novel solutions regarding many of the above mentioned aspects following demands of real applications. On one hand we combine various statistical features to solve different change detection problems, and explore the connections between different 2D image and 3D point cloud based descriptors. On the other hand, we investigate the efficiency of various possible model structures both in terms of scalability and in practical problem solving performance. We will propose new complex and flexible low level inference algorithms between various measured features and prior knowledge. Dealing with higher dimensional data we pay particular attention to reduce the complexity of the model structures, to save computational time and keep the modeling process tractable.

A higher level of visual data interpretation can be based on object level analysis of the scene. Object extraction is a crucial step in several perception applications, starting from remotely sensed data analysis, through optical fabric inspection systems, until video surveillance.

Object detection techniques in the literature follow either a bottom-up, or an inverse (top-down) approach. The straightforward *bottom-up* techniques [78] construct the objects from primitives,

like blobs, edge parts or corners in images. Although these methods can be fast, they may fail if the primitives cannot be reliably detected. We can mention here Hough transform or mathematical morphology based methods [79] as examples, however these approaches show limitations in cases of dense populations with several adjacent objects. To increase robustness, it is common to follow the Hypothesis Generation-Acceptance (HGA) scheme [80, 81]. Here the accuracy of object proposition is not crucial, as false candidates can be eliminated in the verification step. However, objects missed by the generation process cannot be recovered later, which may result in several false negatives. On the other hand, generating too many object hypotheses (e.g. applying exhaustive search) slows down the detection process significantly. Finally, conventional HGA techniques search for separate objects instead of global object configurations, disregarding population-level features such as overlapping, relative alignment, color similarity or spatial distance of the neighboring objects [82].

To overcome the above drawbacks, recent *inverse methods* [83] assign a fitness value to each possible object configuration, and an optimization process attempts to find the configuration with the highest confidence. This way, flexible object appearance models can be adopted, and it is also straightforward to incorporate prior shape information and object interactions. However, this inverse approach needs to perform a computationally expensive search in a high dimensional population space, where local maxima of the fitness function can mislead the optimization.

Using the above terminology, MRFs can also be considered as *inverse* techniques. However staying at pixel level in the graph nodes, we find only very limited options to consider geometrical information [84, 85]. Marked Point Processes (MPP) [83, 86] offer an efficient extension of MRFs, as they work with objects as variables instead of with pixels, considering that the number of variables (i.e. number of objects) is also unknown. MPPs embed prior constraints and data models within the same density, therefore similarly to MRFs, efficient algorithms for model optimization [87, 88, 89] and parameter estimation [90, 91] are available. Recent MPP applications range from 2D [92, 93] and 3D object extraction [9, 94] in various environments, to 1D signal modeling [95] or target tracking [96].

Marked Point Processes have previously been used for various population counting problems, dealing with a large number of objects which have low varieties in shape. MPP models can efficiently handle these situations, through jointly describing individual objects by various data terms, and using information from entity interactions by prescribing the (soft) fulfillment of prior geometric constraints [86]. In this way, one can extract configurations which are composed of similarly shaped and sized entities such as buildings [97], trees [98, 99], birds [87, 88, 94], or boats [90] from remotely sensed data, cell nuclei from medical images [100], galaxies in space applications [93] or people in video surveillance scenarios [101]. While the computational complexity of MPP optimization may mean bottleneck for some applications, various efficient techniques have

been proposed to speed up the energy minimization process, such as the Multiple Birth and Death (MBD) [87] algorithm or the parallel Reversible-Jump Markov Chain Monte Carlo (RJMCMC) sampling process [89].

Although the above applications show clear practical advantages of conventional MPP based solutions, neither the time dimension of the measurements nor the spatial hierarchical decomposition of the scene are addressed in the referred previous works of the literature. Therefore, this thesis presents contributions focusing on temporal and spatial extensions of the original MPP framework, by expansively analyzing the needs and alternative directions for the solutions, and demonstrating the advantages of the improvements in real problem environments.

The temporal dimension appears in two different aspects. The *first problem* is object-level change detection in image pairs, where low level approaches are combined with geometric object extraction by a multi temporal MPP (mMPP) model. The result is an object population, where each object is marked as *unchanged*, *changed*, *new*, or *disappeared* between the selected two time instances, typically based on measurements with several months or years time differences. A *second task* is tracking a moving target across several frames in time sequences of very low quality measurements, such as radar images. For this purpose, a novel Multiframe MPP (F^mMPP) framework is proposed, which simultaneously considers the consistency of the observed data and the fitted objects in the individual images, and also exploits interaction constraints between the object parameters in the consecutive frames of the sequence. Following the Markovian approach, here each target sample may only affect objects in its *neighboring frames* directly, limiting the number of interactions for efficient sequence analysis.

Another major targeted issue is spatial hierarchical content modeling. Classical MPP-based image analysis models [83, 87] focus purely on the object level of the scene. Simple prior interaction constraints such as non-overlapping or parallel alignment are often utilized to refine the accuracy of detection, but in this way only very limited amount of high level structural information can be exploited from the global scenario. In various applications however, investigation of object grouping patterns and the decomposition of objects to smaller parts (i.e. sub-objects) are relevant issues. We propose therefore a hierarchical MPP extension, called the Embedded Marked Point Process (EMPP) model, which encapsulates on one hand a hierarchical description between objects and object parts as a parent-child relationship, and on the other hand it allows corresponding objects to form coherent object groups, by a Bayesian segmentation of the population.

Machine based perception and analysis of the dynamic 3D (*i.e.* 4D, where the 4th dimension is time) environment is nowadays a hot topic in research and engineering, following the quick progress of autonomous driving, security and smart city related applications. While *conventional* electro-optical cameras are still important visual information sources, recently released Lidar range

sensors offer alternative approaches for scene analysis, by directly measuring 3D geometric information from the environment. Using the Lidar technology, the most important limitation is currently a necessary trade-off between the spatial and the temporal resolution of the available sensors, which makes difficult to observe and analyze small details of the scenes in real time. Important research issues are therefore the exploration of new tasks which can be handled by these new sorts of 4D measurements, the adaption of conventional image processing algorithms, structures for voxel-based scene representation and vision related machine learning methodologies to Lidar data, and the fusion of measurements of various Lidar and optical sensors to obtain a more complete scene model. We deal in this thesis with three selected problem families of 4D environment perception. *First*, we propose a new Bayesian approach for person localization and height estimation in a multi-camera system. *Second*, we construct a novel people surveillance framework based on the measurements of a single Rotating Multi-beam (RMB) Lidar sensor, implementing motion detection, moving object separation, tracking, and biometric person identification via Lidar-based gait descriptors. *Third*, we introduce a new workflow with various novel algorithms, for large-scale urban environment analysis based on a car-mounted RMB Lidar, also using a very detailed 3D reference map obtained via laser scanning.

This thesis uses the basic concepts and results of probability theory (e.g. random variables, probability density functions, Bayes rule etc.), and machine learning (neural networks, supervised training strategies) which are supposed to be familiar for the Readers.

The outline of the dissertation is as follows. Chapter 2 presents a short introduction to stochastic image segmentation, object population extraction and machine learning approaches, by introducing the data types, general notations and basic mathematical tools used in the following parts of the thesis. The scientific contributions of the Author are presented in Chapters 3-6. Each of these chapters corresponds to a Thesis Group listed in the Conclusion part in Section 7.2, by giving the background of the selected problems and the main steps of the solution, particularly focusing on the validation of the new scientific results which is performed in experimental ways in most cases. In Chapter 3 novel multi-layer Markovian label fusion models are proposed for two different change detection applications. Chapter 4 deals with multitemporal object level scene analysis for tasks of building change detection in remotely sensed optical image pairs, and moving target tracking in radar image sequences. In Chapter 5 we give a complex multi-level stochastic model for spatial scene decomposition, and demonstrate its usability in three very different application fields. Finally, in Chapter 6 we introduce our above detailed contributions connected to the 4D environment perception topic. A short conclusion and a summary of the new scientific results is given in Chapter 7. For helping the Reader, Appendix A provides a detailed overview on the used abbreviations and notations, while Appendices B-E include some additional figures, tables and pseudo codes connected to the main contributions of the thesis.

Chapter 2

Fundamentals

In this thesis, the various sensor measurements at given time instances are represented either as 2D digital images or as 3D point clouds. Both cases can be completed with a temporal dimension obtaining image or point cloud sequences.

A digital image is defined over a two dimensional pixel lattice S having a finite size $S_W \times S_H$, where $s \in S$ denotes a single pixel. The pixels' observation values represent grayscale or RGB color information, depth values etc. or any descriptors calculated from the raw sensor measurements by spatio-temporal filtering or feature fusion.

A point cloud $\mathcal{L}$ is by definition an unordered set of l points: $\mathcal{L} = \{p_1, \ldots, p_l\}$, where each point, $p \in \mathcal{L}$, is described by its (x, y, z) position coordinates in a 3D Euclidean world coordinate system. Additional parameters, such as intensity, color values or further sensor-specific parameters may also be associated with the points.

Although several different techniques are discussed in the thesis with various goals and model structures, they are strongly connected from the point of view theoretical foundations and methodologies: many of them can be formulated either as *low level segmentation* (or classification) problems or as *object population extraction* tasks (see examples in Fig. 2.1).

Segmentation (or classification) can be formally considered as a labeling task where each local element (pixel of the image or point of the point cloud) gets a label from a J-element label set corresponding to J different segmentation classes. In other words, a J-colored image or point cloud is generated for a given input. Following statistical *inverse* approaches, we should be able to assign a fitness (or probability) value to all the $J^{\#\mathrm{el}}$ possible segmentations[1], based on the current measurements (called *observation*), domain specific knowledge about the classes, and prior constraints, in a way that higher fitness values correspond to semantically better solutions.

By *object population extraction*, we mean the detection of an unknown number of entities from a preliminary defined object library. Here the fitness function needs to characterize any of

[1] $\#\mathrm{el}$ marks here the number of pixels, or points

(a) Region classification

(b) Building extraction

Figure 2.1: Demonstration of a *segmentation* and an *object population extraction* task for aerial images.

the possible entity configurations. The objects are described by geometric shapes such as ellipses or rectangles, while the fitness function evaluates how the independent objects fit the image data and it may also consider pre-defined interaction constraints.

To overcome the course of dimensionality, the fitness functions are usually modularly defined: they can be decomposed into individual subterms, and the domain of each subterm consists only of a few nearby pixels or objects. In this way, if we change locally the segmentation map or the population, we should not re-calculate the whole fitness function, only those subterms, which are affected by the selected entities. This property significantly decreases the computational complexity of iterative optimization techniques [102, 103].

An efficient Bayesian approach can be based on a graph representation, where each node of the graph corresponds to a structural model element, such as a pixel of the image, or an object of the population. We define edges between two nodes, if the corresponding entities influence each other directly, i.e. there is a subterm of the fitness function which depends on both elements. For example, to ensure the spatial smoothness of the segmented images, one can prescribe that the neighboring pixels should have the same labels in the vast majority of cases [104].

Since the seminal work of Geman and Geman [102], Markov Random Fields (MRFs) and their variants such as Mixed Markov models offer powerful tools to ensure contextual classification in image or point set segmentation tasks. Marked Point Process (MPP) models have been introduced in computer vision more recently, as a natural object-level extension of MRFs. In the following part of the chapter we give the formal definitions and algorithmic steps regarding MRF based data segmentation and MPP based object population extraction. The concepts and notations introduced here will be used in the following parts of the thesis.

2.1 Markovian classification models

2.1.1 Markov Random Fields, Gibbs Potentials and Observation Processes

A Markov Random Field (MRF) can be defined over an undirected graph $\mathcal{G} = (\mathcal{V}, \varepsilon)$, where $\mathcal{V} = \{\upsilon_i | i = 1, \ldots N\}$ marks the set of nodes, and ε is the set of edges. Two nodes υ_i and υ_k are *neighbors*, if there is an edge $e_{ik} \in \varepsilon$ connecting them. The set of points which are neighbors of a node υ (i.e. the neighborhood of υ) is denoted by $\mathcal{N}_\upsilon$, while we mark with $\mathcal{N} = \{\mathcal{N}_\upsilon | \upsilon \in \mathcal{V}\}$ the neighborhood system of the graph.

A classification problem can be interpreted as a labeling task over the nodes. Using a finite label set $\Lambda = \{l_1, l_2, \ldots, l_J\}$, we assign a unique label $\varsigma(\upsilon) \in \Lambda$ to each node $\upsilon \in \mathcal{V}$. We mean by a *global labeling* ϖ the enumeration of the nodes with their corresponding labels:

$$\varpi = \{\ [\upsilon, \varsigma(\upsilon)] \mid \forall \upsilon \in \mathcal{V}\ \}. \tag{2.1}$$

Let us denote by Υ the (finite) set of all the possible global labellings ($\varpi \in \Upsilon$).

In some cases, instead of a global labeling, we need to deal with the labeling of a given subgraph. The subconfiguration of ϖ with respect a subset $X \subseteq \mathcal{V}$ is denoted by $\varpi_X = \{\ [\upsilon, \varsigma(\upsilon)] \mid \forall \upsilon \in X\ \}$.

In the next step, we define Markov Random Fields. As usual, Markov property means here that the label of a given node depends only on its neighbors directly.

Definition 1 ***(Markov Random Field)*** *$\mathcal{X}$ is a Markov Random Field (MRF), with respect to a graph $\mathcal{G}$, if the following two conditions hold:*

- *for all $\varpi \in \Upsilon$; $P(\mathcal{X} = \varpi) > 0$*
- *for every $\upsilon \in \mathcal{V}$ and $\varpi \in \Upsilon$: $P(\varsigma(\upsilon) \mid \varpi_{\mathcal{V}\setminus\{\upsilon\}}) = P(\varsigma(\upsilon) \mid \varpi_{\mathcal{N}_\upsilon})$.*

Discussion about MRFs is most convenient by defining the neighborhood system via the *cliques* of the graph. A subset $C \subseteq \mathcal{V}$ is a clique if every pair of distinct nodes in C are neighbors. $\mathcal{C}$ denotes a set of cliques.

To characterize the *fitness* of the different global labellings, a Gibbs measure is defined on Υ. Let V be a potential function which assigns a real number $V_X(\varpi)$ to the subconfiguration ϖ_X. V defines an energy $U(\varpi)$ on Υ by

$$U(\varpi) = \sum_{X \in 2^{\mathcal{V}}} V_X(\varpi). \tag{2.2}$$

where $2^{\mathcal{V}}$ denotes the set of the subsets of $\mathcal{V}$.

Definition 2 ***(Gibbs distribution)*** *A Gibbs distribution is a probability measure π on Υ with the following representation:*

$$\pi(\varpi) = \frac{1}{Z} \exp\Big(- U(\varpi)\Big) \tag{2.3}$$

where Z is a normalizing constant or partition function:

$$Z = \sum_{\varpi \in \Upsilon} \exp\big(- U(\varpi)\big). \tag{2.4}$$

If $V_X(\varpi) = 0$ whenever $X \notin \mathcal{C}$, then V is called a nearest neighbor potential.

The following theorem is the principle of most MRF applications in computer vision [102]:

Theorem 1 ***(Hammersley-Clifford)*** *$\mathcal{X}$ is an MRF with respect to the neighborhood system $\mathcal{N}$ if and only if $\pi(\varpi) = P(\mathcal{X} = \varpi)$ is a Gibbs distribution with nearest neighbor Gibbs potential V, that is*

$$\pi(\varpi) = \frac{1}{Z} \exp\left(- \sum_{C \in \mathcal{C}} V_C(\varpi) \right) \tag{2.5}$$

We mean by *observation* arbitrary measurements from real world processes (such as image sources) assigned to the nodes of the graph. In image processing, usually the pixels' color values or simple textural responses are used, but any other local features can also be calculated. In general, we only prescribe that the observation process assigns a D dimensional real vector, $f(\upsilon) \in \mathbb{R}^D$, to selected graph nodes. The global observation over the graph is marked by

$$\mathcal{F} = \{\, [\upsilon, f(\upsilon)] \mid \forall \upsilon \in O \,\} \text{ where } O \subseteq \mathcal{V}. \tag{2.6}$$

MRF based classification models use two assumptions. *First*, each class label $l_i \in \Lambda$ corresponds to a random process, which generates the observation value $f(\upsilon)$ at υ according to a locally specified probability density function *(pdf)*, $p_{\upsilon,i}(\lambda) = P(f(\upsilon) = \lambda | \varsigma(\upsilon) = l_i)$. *Second*, local observations are conditionally independent, given the global labeling:

$$P(\mathcal{F}|\varpi) = \prod_{\upsilon \in O} P\left(f(\upsilon)|\varsigma(\upsilon)\right). \tag{2.7}$$

2.1.2 Bayesian labeling approach and the Potts model

Let $\mathcal{X}$ be an MRF on graph $\mathcal{G} = (\mathcal{V}, \varepsilon)$, with (a priori) clique potentials $\{V_C(\varpi) \mid C \in \mathcal{C}\}$. Consider an observation process $\mathcal{F}$ on $\mathcal{G}$. The goal is to find the labeling $\widehat{\varpi}$, which is the maximum a posteriori (MAP) estimate, i.e. the labeling with the highest probability given $\mathcal{F}$:

$$\widehat{\varpi} = \underset{\varpi \in \Upsilon}{\operatorname{argmax}}\, P(\varpi|\mathcal{F}). \tag{2.8}$$

Following Bayes' rule and eq. (2.7),

$$P(\varpi|\mathcal{F}) = \frac{P(\mathcal{F}|\varpi)P(\varpi)}{P(\mathcal{F})} = \frac{1}{P(\mathcal{F})}\left[\prod_{\upsilon\in O} P\left(f(\upsilon)|\varsigma(\upsilon)\right)\right] P(\varpi) \tag{2.9}$$

Based on the Hammersley-Clifford theorem, $P(\varpi)$ follows a Gibbs distribution:

$$P(\varpi) = \pi(\varpi) = \frac{1}{Z}\exp\left(-\sum_{C\in\mathcal{C}} V_C(\varpi)\right) \tag{2.10}$$

while $P(\mathcal{F})$ and Z (in the Gibbs distribution) are independent of the current value of ϖ. Using also the monotonicity of the logarithm function and equations (2.8), (2.9), (2.10), the optimal global labeling can be written into the following form:

$$\widehat{\varpi} = \underset{\varpi\in\Upsilon}{\operatorname{argmin}}\left\{\sum_{\upsilon\in O} -\log P\left(f(\upsilon)|\varsigma(\upsilon)\right) + \sum_{C\in\mathcal{C}} V_C(\varpi)\right\}. \tag{2.11}$$

Note that due to the conditional independence of the observations at the different nodes, the fact that the prior field $\pi(\varpi)$ is an MRF implies that the $\pi(\varpi|\mathcal{F})$ posterior field is also an MRF. In this case the $-\log P\left(f(\upsilon)|\varsigma(\upsilon)\right)$ quantity can be considered as the potential of a *singleton clique* $\{\upsilon\}$.

2.1.3 MRF based image segmentation

A widely used implementation of the above Bayesian labeling framework for image segmentation is based on the Potts model [104]. Assume that the problem is defined over the 2D lattice S and we have a measurement vector $f(s) \in \mathbb{R}^D$ at each pixel s. The goal is to segment the input lattice with J pixel clusters corresponding to J random processes $(l_1, \ldots, l_J)$, where the clusters of the pixels are consistent with the local measurements, and the segmentation is *smooth*, i.e. pixels having the same cluster form connected regions. Here by the definition of $\mathcal{G}$, we assign to each pixel of the input lattice a unique node of the graph. One can simply use a *first ordered* neighborhood, were each pixel has four neighbors. In this case, the cliques of the graph are singletons or doubletons as shown in Fig. 2.2. As a consequence, the prior term $\pi(\varpi) = P(\varpi)$ of the MRF energy function is defined by the doubleton clique potentials. According to the Potts model, the prior probability term is responsible for getting smooth connected components in the segmented image, so that we give penalty terms to each neighboring pair of nodes whose labels are different. For any $r, \upsilon \in \mathcal{V}$ node pairs, which fulfill $\upsilon \in \mathcal{N}_r$, $\{r, \upsilon\} \in \mathcal{C}$ is a clique of the graph, with the potential:

$$V_{\{r,\upsilon\}}(\varpi) = \begin{cases} -\delta & \text{if } \varsigma(r) = \varsigma(\upsilon) \\ +\delta & \text{if } \varsigma(r) \neq \varsigma(\upsilon) \end{cases} \tag{2.12}$$

where $\delta \geq 0$ is a constant.

A sample MRF based segmentation result, with the demonstration of the role of the Potts smoothing term, is shown in Fig. 2.3.

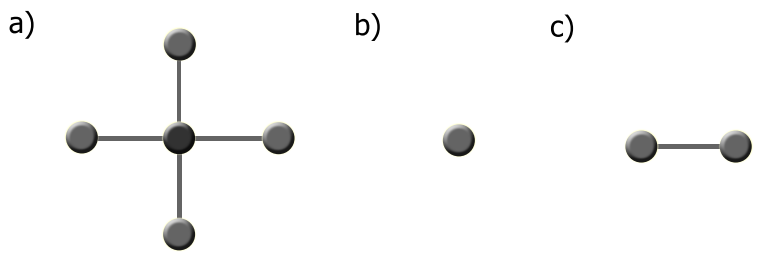
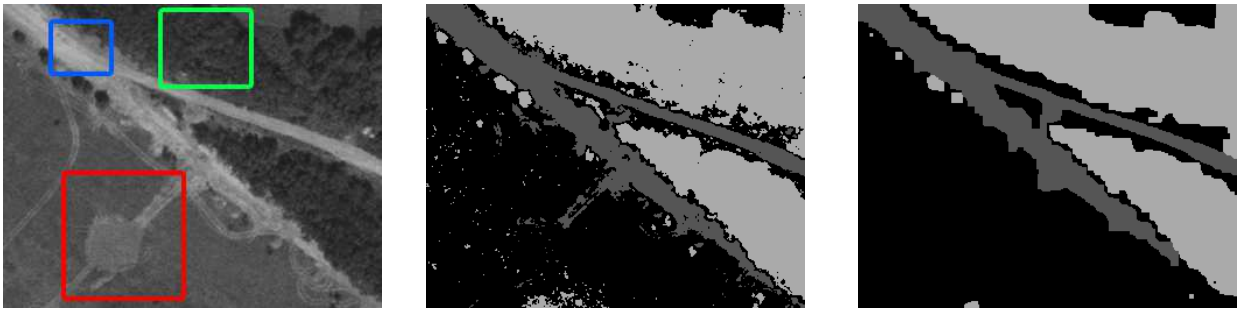


Figure 2.2: Illustration of simple connections in MRFs: (a) first ordered neighborhood of a selected node on the lattice, (b) 'singleton' clique, (c) doubleton clique



(a) (b) (c)

Figure 2.3: Demonstration of MRF based supervised image segmentation with three classes: (a) input image with the training regions, (b) pixel-by-pixel segmentation without using node interactions, (c) result of the Potts model with MMD optimization [53]

2.1.4 MRF Optimization

In applications using the MRF models, the quality of the classification depends both on the appropriate probabilistic model of the classes, and on the optimization technique which finds a good global labeling with respect to eq. (2.11). The latter factor is a key issue, since finding the global optimum is NP hard [105]. On the other hand, stochastic optimizers using simulated annealing (SA) [102, 103] and graph cut techniques [105] have proved to be practically efficient offering a ground to validate different energy models.

The results shown in the following chapters have been generated by either the deterministic Modified Metropolis (MMD) [106, 107] relaxation algorithm or by the fast graph-cut based optimization technique [105]. Detailed overviews on the various optimization approaches, and tutorials on MRF based image segmentation can be found in the *Ph.D. dissertation* of the Author [53], and in several books and monographs dealing with the topic [73, 77].

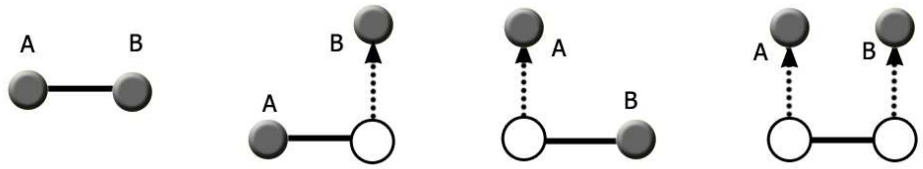


Figure 2.4: Possible interactions in mixed Markov models. Four different configurations, where A and B regular nodes may directly interact. Empty circles mark address nodes, continuous lines are edges, dotted arrows denote address pointers.

2.1.5 Mixed Markov Models

Mixed Markov models have been originally proposed for gene regulatory network analysis [75], and extend the modeling capabilities of Markov random fields: besides prior *static* connections, they enable using observation-dependent *dynamic* links between the processing nodes. This property allows encoding interactions that occur only in a certain context and are absent in all others. A mixed Markov model – similarly to a conventional MRF – is defined over a graph $\mathcal{G} = (\mathcal{V}, \varepsilon)$, where $\mathcal{V}$ and ε denote again the sets of nodes and edges, respectively. A label, i.e. a random variable $\varsigma(v)$, is assigned to each node $v \in \mathcal{V}$ as well, and the node labels over the graph determine a *global labeling* ϖ as defined by Formula (2.1).

However in mixed Markov models two types of nodes are discriminated: $\mathcal{V}_R$ contains *regular nodes* and $\mathcal{V}_A$ is the set of *address nodes* ($\mathcal{V} = \mathcal{V}_R \cup \mathcal{V}_A,\ \mathcal{V}_R \cap \mathcal{V}_A = \emptyset$). Regular nodes $r \in \mathcal{V}_R$ have the same roles as nodes in MRFs: the corresponding variable $\varsigma(r)$ will encode a segmentation label getting values from a finite, application dependent label set. On the other hand address nodes provide configurable links in the graph by creating pointers to other (regular) nodes. Thus for a given address node $a \in \mathcal{V}_A$, the domain of its 'label' $\varsigma(a)$ is the set $\mathcal{V}_R \cup \{\text{nil}\}$. In the case of $\varsigma(a) \neq \text{nil}$, let us denote by $\varsigma^\star(a)$ the label of the regular node addressed by a:

$$\varsigma^\star(a) := \varsigma(\varsigma(a)). \tag{2.13}$$

There is no restriction on the graph topology: edges can link any two nodes. The edges define the set of cliques of $\mathcal{G}$, which is denoted again by $\mathcal{C}$.

In a given configuration, two regular nodes may interact directly if they are connected by a static edge or by a chain of a static edge and dynamic address pointers: four typical configurations of connection are demonstrated in Fig. 2.4. More specifically, with notation for each clique $C \in \mathcal{C}$: $\varsigma_C = \{\varsigma(v) | v \in C\}$ and $\varsigma_C^A = \{\varsigma^\star(a) \big| a \in \mathcal{V}_A \cap C, \varsigma(a) \neq \text{nil}\}$ the prior probability of a given global labeling ϖ is given by:

$$P(\varpi) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \exp\Big(- V_C\left(\varsigma_C, \varsigma_C^A\right) \Big) \tag{2.14}$$

(a) (b) (c)

Figure 2.5: Marked Point Process example: (a) building population as a realization of a point process (b) rectangle model of a selected building, (c) parameters of the marked object [10]

where V_C is a $\mathcal{C} \rightarrow \mathbb{R}$ *clique potential function*, which has a 'low' value if the labels within the set $\varsigma_C \cup \varsigma_C^A$ are semantically consistent, while V_C is 'high' otherwise. Scalar $Z = \sum_{\varpi} P(\varpi)$ is again a normalizing constant, which could be calculated over all the possible global labelings. Note that a detailed analysis of analytical and computational properties of mixed Markov models can be found in [75], which confirms the efficiency of the approach in probabilistic inference.

2.2 Object population extraction with Marked Point Processes

Similarly to Markov Random Fields, the Marked Point Process (MPP) methods use a graph-based representation for semantic content modeling. However, in MPPs the graph nodes are associated with geometric objects instead of low level pixels or point cloud elements. In this way an MPP model enables to characterize whole populations instead of individual objects, through exploiting information from entity interactions. Following the classical Markovian approach, each object may only affect its *neighbors* directly. This property limits the number of interactions in the population and results in a compact description of the global scene, which can be analyzed efficiently.

For easier discussion, in this chapter we introduce MPP models purely over 2D pixel lattices, dealing with 2D objects. While most of the object detection tasks discussed in this thesis are handled indeed as 2D pattern recognition problems, we will show in later chapters, that the model extension to 2.5D or 3D (spatial) scenes is quite straightforward.

2.2.1 Definition of Marked Point Processes

In statistics, a random process is called *point process*, if it can generate a set of isolated points either in space or time, or in even more general spaces. In this thesis we will mainly use a discrete *2D point process*, whose realization is a set of an arbitrary number of points over a pixel lattice S:

$$\overline{o} = \{o_1, o_2, \ldots, o_n\}, \quad n \in \{0, 1, 2, \ldots\}, \; \forall i : \; o_i \in S. \tag{2.15}$$

A sample task for using point processes in image processing is detecting buildings in aerial images, as shown in Fig. 2.5(a), where each point corresponds to a building center. However, modeling our objects with point-wise entities is often an insufficient abstraction. For example, in high resolution aerial photos building shapes can often be efficiently approximated by rectangles (Fig. 2.5(b)). To include object geometry in the model, we assign markers to the points. As shown in Fig. 2.5(c), a rectangle can be defined by the center point $o \in S$, the orientation $\theta \in [-90°, +90°]$ and the perpendicular side lengths e_L and e_l. In this case the marker is a 3D parameter vector (θ, e_L, e_l).

Taking a general case, let us denote by u an object candidate of the scene whose imaged shape over lattice S is represented by a plane figure from a preliminary fixed shape library. In this thesis, ellipses, rectangles and isosceles triangles are used. We will model each marked object by its reference point o, the global orientation θ and further shape dependent parameters such as major and minor axes for ellipses, the perpendicular side lengths for rectangles, and a side-altitude pair for triangles. With denoting by $\mathcal{P}$ the domain of the markers, the $\mathcal{H}$ parameter space of the individual objects (i.e. $u \in \mathcal{H}$) is obtained as $\mathcal{H} = S \times \mathcal{P}$.

A configuration of an MPP model, denoted by ω, is a population of marked objects:

$$\omega = \{u_1, \ldots, u_n\}, \;\; \forall i : u_i \in \mathcal{H}, \tag{2.16}$$

where the number of objects, n, is an arbitrary integer, which is initially unknown in population extraction tasks. Consequently, the object configuration space, Ω, has the following form:

$$\Omega = \bigcup_{n=0}^{\infty} \Omega_n, \quad \Omega_n = \big\{\{u_1, \ldots, u_n\} \subset \mathcal{H}^n\big\}. \tag{2.17}$$

Next, we define a $\sim$ neighborhood relation between the objects of a given ω configuration. For example, we can prescribe for objects $u, v \in \omega$, that $u \sim v$ iff the distance between the object centers is lower than a predefined threshold. The neighborhood of object u in ω is:

$$\mathcal{N}_u(\omega) = \{v \in \omega | u \sim v\}. \tag{2.18}$$

2.2.2 MPP energy functions

Object populations in MPP models are evaluated by simultaneously considering the input measurements (e.g. images), and prior application specific constraints about object geometry and interactions. Let us denote by $\mathcal{F}$ the union of all image features derived from the input data. For characterizing a given ω configuration based on $\mathcal{F}$, we introduce a non-homogenous data-dependent Gibbs distribution (see eq.(2.3)) on the population space:

$$P_{\mathcal{F}}(\omega) = P(\omega|\mathcal{F}) = \frac{1}{Z} \cdot \exp\Big(-\Phi_{\mathcal{F}}(\omega)\Big) \tag{2.19}$$

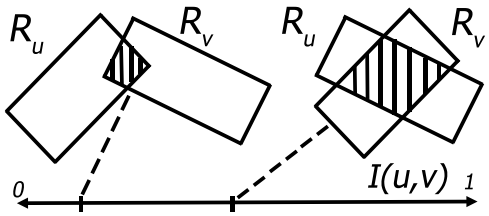


Figure 2.6: Calculation of the $I(u, v)$ interaction potentials: intersections of rectangles are denoted by striped areas

with a Z normalizing constant:

$$Z = \sum_{\omega \in \Omega} \exp\big(-\Phi_{\mathcal{F}}(\omega)\big), \tag{2.20}$$

$\Phi_{\mathcal{F}}(\omega)$ is called the configuration energy. Following the energy decomposition approach discussed earlier by MRFs (eq. (2.5)), we obtain $\Phi_{\mathcal{F}}(\omega)$ as the sum of simple components, which can be calculated by considering small subconfigurations only. More specifically, we distinguish unary (or singleton) terms (A) defined on individual objects, and *I*nteraction terms ($I(u, v)$), concerning neighboring objects:

$$\Phi_{\mathcal{F}}(\omega) = \sum_{u \in \omega} A(u) + \gamma \cdot \sum_{\substack{u,v \in \omega \\ u \sim v}} I(u, v) \tag{2.21}$$

$\gamma > 0$ is a weighting factor between the unary and interaction terms, and it should be calibrated in each application in a case-by-case basis. In general, both the $A(u)$ and $I(u, v)$ terms may depend on the $\mathcal{F}$ observation. However, it is a frequent strategy that only the unary terms depend on $\mathcal{F}$, so that they evaluate the object candidates as a function of the local image data. On the other hand, the $I(u, v)$ components may implement prior geometric constraints, such as neighboring objects should not overlap, or they should have similar orientation. Denoting by $R_u \subset S$ the set of image pixels covered by the geometric figure of object u, a simple interaction term penalizing object intersection can be calculated as:

$$I(u, v) = \frac{\#(R_u \cap R_v)}{\#(R_u \cup R_v)} \tag{2.22}$$

where $\#$ denotes the set cardinality. (See also Fig. 2.6.)

In the following, we will only use the subscript $\mathcal{F}$, when we want to particularly emphasize that a given MPP energy term depends on the measurement data (eg. $A_{\mathcal{F}}(u)$, $\Phi_{\mathcal{F}}(\omega)$). In several clear situations the subscript notation will be omitted to preserve the simplicity of formalism.

2.2.3 MPP optimization

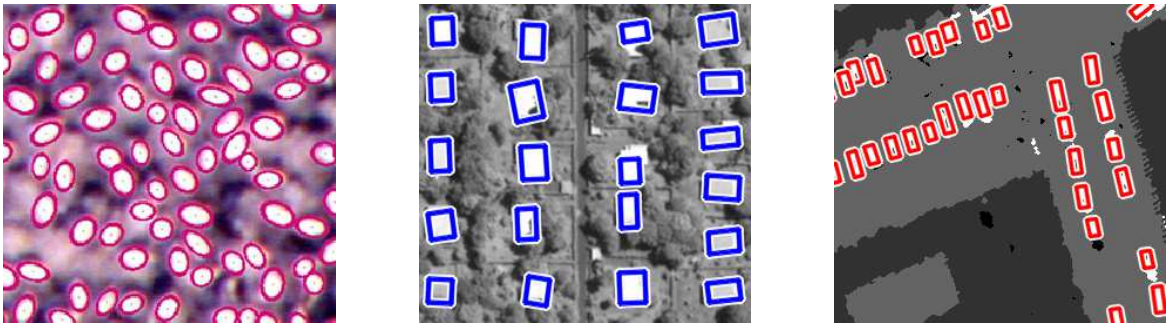
The optimal object population $\widehat{\omega}$ in an MPP model can be taken as the MAP configuration estimate:

$$\widehat{\omega} = \underset{\omega \in \Omega}{\operatorname{argmax}} P_{\mathcal{F}}(\omega) = \underset{\omega \in \Omega}{\operatorname{argmin}} \Phi_{\mathcal{F}}(\omega). \tag{2.23}$$

However, finding $\widehat{\omega}$ needs to perform an efficient search in the high dimension population space with a non-convex energy function. Ensuring high quality object configurations by algorithms with feasible computation complexity is crucial in several applications, therefore we can find an extensive bibliography of MPP energy minimization techniques. Most previous approaches use the iterative Reversible Jump Markov Chain Monte Carlo (RJMCMC) scheme [108, 109], where each iteration consists in perturbing one or a couple of objects using various kernels such as birth, death, translation, rotation or dilation. Here experiments show that the rejection rate, especially for the birth move, may induce a heavy computation time. Besides, one should be very careful when decreasing the temperature, because at low temperature, it is difficult to add objects to the population.

A recent alternative approach, called the Multiple Birth and Death Dynamics technique (MBD) [87] attempts to overcome several ones from the above mentioned limitations. Unlike following a discrete jump-diffusion scheme like in RJMCMC, the MBD optimization method defines a continuous time stochastic evolution of the object population, which aims to converge to the optimal configuration. The evolution under consideration is a birth-and-death equilibrium dynamics on the configuration space, embedded into a Simulated Annealing (SA) process, where the temperature of the system tends to zero in time. The final step is the discretization of this non-stationary dynamics: the resulting discrete process is a non-homogeneous Markov chain with transition probabilities depending on the temperature, energy function and discretization step. In practice, the MBD algorithm evolves the population of objects by alternating purely stochastic object generation (*birth*) and removal (*death*) steps in a SA framework. In contrast to the above RJMCMC implementations, each birth step of MBD consists of adding *several* random objects to the current configuration, which is allowed due to the discretization trick. Using MBD, there is no rejection during the birth step, therefore high energetic objects can still be added independently of the temperature parameter. Thus the final result is much less sensitive to the tuning of the SA temperature decreasing process, which can be achieved faster. Due to these properties, in selected remote sensing tasks (bird and tree detection) [87] the optimization with MBD proved to be around ten times faster than RJMCMC with similar quality results. On the other hand, we note that parallel sampling in MBD implementations is less straightforward than regarding the RJMCMC relaxation [89].

In the thesis, we will propose different structural modifications of the Multiple Birth and Death Dynamic (MBD) adopted to our addressed problems. For a deeper understanding of this approach, we introduce here the steps of the basic MBD algorithm [87]:



(a) Flamingos © INRIA (b) Building detection (c) Vehicle detection

Figure 2.7: Selected examples of population extraction with MPP models: (a) flamingo detection in aerial images [87], (b) building extraction in satellite photos [10], (c) vehicle detection from Lidar data [34]

1. *Initialization:* calculate a $P_b() : S \to \mathbb{R}$ birth map using the $\mathcal{F}$ input data, which assigns to each pixel s a pseudo probability value $P_b(s)$ estimating how likely s is an object center.

2. *Main program:* initialize the inverse temperature parameter $\beta = \beta_0$ and the discretization step $\delta = \delta_0$ and alternate birth and death steps.

   (a) *Birth step*: for each pixel $s \in S$, if there is no object with center s in the current configuration ω, choose birth with probability $\delta P_b(s)$.

   If birth is chosen in s:

   - generate a new object u with center s
   - set the object parameters (marks of u) randomly based on prior knowledge
   - add u to the current configuration ω.

   (b) *Death step*: Consider the configuration of objects $\omega = \{u_1, \ldots, u_n\}$ and sort it from the highest to the lowes value of the unary (data) term $\varphi_Y(u)$. For each object u taken in this order compute the death rate as follows:

   $$d_\omega(u) = \frac{\delta a_\omega(u)}{1 + \delta a_\omega(u)}, \text{ where } a_\omega(u) = \exp\left[-\beta\Big(\Phi_{\mathcal{F}}(\omega/\{u\}) - \Phi_{\mathcal{F}}(\omega)\Big)\right]$$

   and kill u with probability $d_\omega(u)$

3. *Convergence test*: if the process has not converged, increase the inverse temperature β and decrease the discretization step δ by a geometric scheme and go back to the birth step. The convergence is obtained when all the objects added during the birth step, and only these ones, have been killed during the death step.

Selected state-of-the-art results for MPP based object population extraction in different applications using different input sources are shown in Fig. 2.7. Examples (b) and (c) are directly related to the thesis.

2.3 Advanced machine learning techniques

The previously discussed Bayesian image classification techniques can be efficiently applied, if either a color/texture based statistical description can specify the semantically corresponding regions (see MRFs), or strong geometric constraints can be adapted for object shape description and object population modeling (MPPs).

In various situations, for example in semantic urban scene segmentation, detection of objects with diverse elastic shapes (e.g. pedestrians, various types of vehicles), or biometric identification based on visual features, such assumptions cannot be set, and neural network (NN) based solutions are often taken as first options. While in MRF/MPP models we directly involve our prior knowledge (such as geometric features) in the modeling process, in NN based methods, the information used for classification should be entirely extracted from the training data, thus the qualitative and quantitative parameters of the training dataset are critical factors. Another crucial issue of NN recognizers is the efficient feature selection. Conventional NNs used handcrafted features specified for each problem separately, thus the feature engineering step was a significant part of algorithm development. This tendency changed by introducing the feature learning strategies of deep neural networks (DNNs), which has grabbed a very intensive focus of computer vision research on machine learning approaches in the recent years [111]. Apart from feature learning, the main contribution of DNNs is that they can also learn strong contextual dependencies from training samples, leading us close to a human-like holistic scene interpretation. On the other hand, while some DNN-based attempts on population counting [112] or object tracking [113] have already been proposed, their superiority versus probabilistic or geometric approaches have not yet been thoroughly demonstrated in these domains.

This thesis does not provide research results on generally improving deep learning methodologies, but as we detail in Chapter 6, we utilize the combination of existing DNN architectures and learning strategies in multiple occasions for solving novel environment perception tasks, including Lidar-based person identification, 3D object recognition and 3D point cloud scene segmentation.

2.4 Methodological contributions of the thesis

Although Markov Random Field (MRF) and Marked Point Process (MPP) models provide established tools for classification and population modeling tasks, they face a couple of limitations, which are disadvantageous in various real work tasks.

In MRF based segmentation models, the integration of multiple information sources is a key issue. Earlier proposed feature fusion approaches, such as observation modeling by multinomial feature distributions, or using simple pixel-by-pixel operations on various label maps, often yield

insufficient performance. In Chapter 3, we propose novel Markovian label fusion models, which enable flexible integration of various observation based and prior knowledge based descriptors in a modular framework. We also introduce a multi-layer Mixed Markov model, which exploits the probabilistic connection modeling capabilities of Mixed Markov models in the multi-layer segmentation process.

The conventional MPP models are extended in this thesis both regarding the temporal and the spatial dimensions. In Chapter 4 we introduce multitemporal MPP frameworks dealing with object level change detection and moving target tracking tasks. From a technical point of view, this extension needs the definition of various data-based or prior interaction terms between object examples from different time layers, apart from the usual intra-layer constraints of eq. (2.22). Regarding spatial scene content decomposition, in Chapter 5 we propose an Embedded MPP model consisting of three hierarchical levels, namely object groups, super objects and object parts. The super (or *parent*) objects play a similar role as regular objects in MPP models, while the object parts (or *child* objects) are also marked objects with a predefined set of possible geometric attributes, and they are connected to the parents through additional markers. On the other hand, the object groups are interpreted as sub-populations, which may contain any number of (parent) objects, and various local geometric constraints can be prescribed for the included members.

Another key point in MPP models is the probabilistic approach for object proposal. In several previous MPP applications [108], the generation of object candidates followed prior (e.g. Poisson) distributions. On the contrary, we apply a data driven birth process to accelerate the convergence of MBD, proposing relevant objects with higher probability based on various image features. In addition, we calculate not only a probability map for the object centers, but also estimate the expected object appearances through low-level descriptors. This approach uses a similar idea to the Data Driven MCMC scheme of image segmentation [110]. However, while in [110] the *importance proposal probabilities* of the moves are used by a jump-diffusion process, we should embed the data driven exploration steps into the MBD framework.

Chapter 6 presents various results from the field of 4D environment perception. For new imaging sensors or sensor configurations, such as the rotating multi-beam Lidar, or the up-to-date high resolution multi-camera systems, even the possible application areas are not completely explored yet. Therefore, several recently published techniques rely on ad-hoc and heuristic methodological approaches. In this thesis, we take the advantage of the established MRF, MPP model concepts fused with various up-to-date machine learning techniques to improve the automatic detection performance under realistic outdoor circumstances.

Chapter 3

Multi-layer label fusion models

In this chapter, new multi-layer Bayesian label fusion models are proposed for two different change detection problems in remotely sensed images.

First a probabilistic model is proposed for automatic change detection on airborne images captured by moving cameras. To ensure robustness, an unsupervised coarse matching is used instead of a precise image registration. The challenge of the proposed model is to eliminate the registration errors, noise and the parallax artifacts caused by the static objects having considerable height (buildings, trees, walls etc.) from the difference image. The background membership of a given image point is described through two different features, and a novel three-layer Markov Random Field (MRF) model is introduced to ensure connected homogeneous regions in the segmented image.

Second, we introduce a Bayesian model, called the Conditional Mixed Markov model (CXM), for detecting relevant changes in registered aerial image pairs taken with the time differences of several years and in different seasonal conditions. The CXM model is a combination of a mixed Markov model and a conditionally independent random field of signals. The new approach integrates global intensity statistics with local correlation and contrast features. A global energy optimization process ensures simultaneously optimal local feature selection and smooth, observation-consistent segmentation. Validation is given on real aerial image sets provided by the Hungarian Institute of Geodesy, Cartography and Remote Sensing and Google Earth.

3.1 Label fusion models in computer vision

As emphasized in Chapter 1, selecting an appropriate model structure is a critical issue for Bayesian image segmentation. Although since Geman and Geman's paper from 1984 [102] extensive research has been conducted on image processing applications of Markov Random Fields (MRF), new focus areas emerged in the 2000's and 2010's due to the evolution of imaging sensor technologies, providing new measurement modalities and enhanced image qualities. The technological progression demanded the development of various new feature fusion approaches within the MRF framework. As notable proposed solutions, we can mention here MRFs using multinomial feature distributions [114, 115], multi-layer MRF models with feature driven inter layer interactions [116], or the fusion MRF [117]. Meanwhile, issues of incorporating novel types of prior information and inference rules in MRFs have been less widely explored in the literature. For this reason, we accomplished research in this direction, focusing on two different pixel level change detection problems:

- **Task 1:** Moving object detection in image pairs captured by moving aerial vehicles with a few seconds of time differences. The task needs an efficient combination of image registration for camera motion compensation and frame differencing (see Fig. 3.1(a)). Registration errors and parallax effects caused by 3D scene structures are modeled as noise components, and a statistical approach is developed to eliminate the undesired distortions form the change mask.

- **Task 2:** Detecting relevant changes in registered aerial images captured with time differences of several months or years. Even staying at a low level (region based) model, this task needs a more sophisticated approach than simple pixel value differencing, since due to seasonal changes or altered illumination, the appearance of the corresponding unchanged areas may also be significantly different. A new region based change detection model is presented, which locally estimates the efficient discriminating features between 'changed' and 'unchanged' image regions (Fig. 3.1(b)).

From a methodological point of view, we present in this chapter four main contributions: *First*, we construct new multi-layer label fusion model structures, which implement flexible integration of various (sub-)segmentation results, with keeping the advantages of the established MRF modeling approach. *Second*, using the Mixed Markovian concept, we introduce dynamic graph structures into the multi-layer framework to extend its modeling capabilities. *Third*, we work out efficient optimization methods for new the multi-layer models. *Fourth*, we give an extensive review and quantitative comparison results of multi-layer models.

(a) Object motion detection with camera motion compensation

(b) Long term change detection

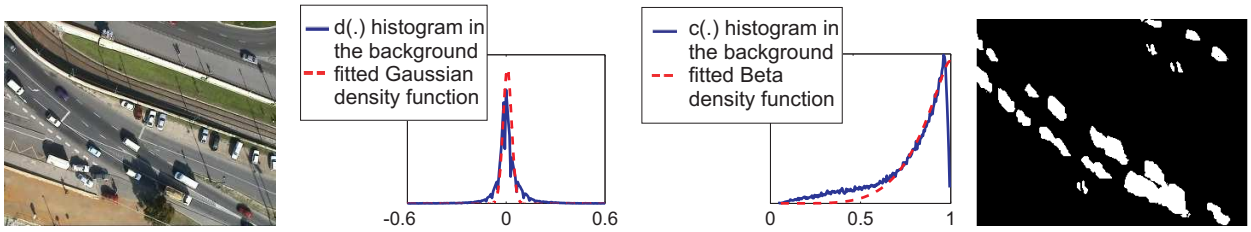
Figure 3.1: Demonstration of the addressed object motion detection and long term change detection problems

In this thesis we mainly focus on introducing the novel methodological issues, while the presentation of the application specific model components serve to demonstrate the main motivation of the developments, and help the Reader in better understanding the models. Note that the corresponding background publications of the Author [12, 13] present detailed introduction and state-of-the-art review regarding the application environments of *task 1* and *task 2*, respectively, with various additional qualitative and quantitative experiments comparing the proposed models to concurrent change detection approaches.

3.2 A label fusion model for object motion detection

As a first example, we focus on the object motion detection problem having two partially overlapped images which were taken by moving airborne vehicles above urban roads with a few seconds time difference [46]. Denote by G_1 and G_2 the two input images above the same pixel lattice S. The gray value of a given pixel $s \in S$ is $g_1(s)$ in the first image and $g_2(s)$ in the second one. Formally, we consider frame differencing as a pixel labeling task with two segmentation classes: foreground (fg) and background (bg). Pixel s belongs to the foreground, if the 3D scene point, which is projected to pixel s in the first frame (G_1), changes its position in the scene's (3D) world coordinate system or is covered by a moving object by the time taking the second image (G_2). Otherwise, pixel s belongs to the background.

Assuming that the observed scene consists of an approximately planar ground region with various static and dynamic 3D urban objects (such as vehicles, walls, trees, short building segments),



(a) First image: G_1 (c) $f_d(.)$ feature statistics (e) $f_c(.)$ feature statistics (g) Ground Truth

(b) Registered second image: $\widetilde{G}_2$ (d) Segmented image D based on $f_d(.)$ (f) Segmented image C based on $f_c(.)$ (h) AND operation on D and C images

Figure 3.2: Feature selection in the multi-layer MRF model. Notations are given in the text of Section 3.2.1.

a 2D similarity transform provided by the Fourier shift-theorem based method [118] can be used for a coarse estimation of the global transform between the images due to camera motion [54]. In the following, the registered second frame is denoted by $\widetilde{G}_2$, and its pixel values by $\{\tilde{g}_2(s)\}$.

3.2.1 Feature selection

Our next task is to define local features at each pixel $s \in S$ which give us information for classifying s as foreground or background point. Thereafter, taking a probabilistic approach, we consider the classes as random processes generating the selected features according to different distributions. The feature selection is shown in Fig. 3.2. The first feature is the gray level difference of the corresponding pixels in $\widetilde{G}_2$ and G_1 respectively: $f_d(s) = \tilde{g}_2(s) - g_1(s)$. As shown in Fig. 3.2(c), the occurring $f_d(.)$ feature values in the background can be statistically characterized by a Gaussian distribution with a given mean value μ (i.e. global intensity offset between the images) and deviation σ (uncertainty due to camera noise and registration errors), while any $f_d(s)$ value may occur in the foreground, hence the foreground class is modeled by a uniform density. Next, we demonstrate the limitations of this feature. After supervised estimation of the distribution parameters, we derive the D image in Fig. 3.2(d) as the maximum likelihood estimate: the label of s is $\mathrm{argmax}_{\psi\in\{\mathrm{fg,bg}\}} P(f_d(s)|\psi)$. We can observe here several false positive foreground points, mainly near to the boundaries of static 3D field objects.

For the above reasons, we introduce a second feature $f_c(s)$, that is obtained by calculating normalized cross correlation between the rectangular pixel neighborhoods $W_1(s)$ in G_1 and $W_2(s + o_s)$ in $\widetilde{G}_2$ for different o_s offset values within an l sided search window and take $f_c(s) =$

$\max_{o_s}\ \mathrm{Corr}\{W_1(s), W_2(s+o_s)\}$. As shown in Fig. 3.2(e), $f_c(s)$ values in the background, can be approximated by a Beta density function [119]: $P(f_c(s)|\mathrm{bg}) = B(f_c(s), \alpha, \beta)$. The foreground class will be described again by a uniform probability $P(f_c(s)|\mathrm{fg})$ with a_c and b_c parameters.

We see in Fig. 3.2(f) [C image] that the $f_c(.)$ descriptor alone causes also poor result. However, unlike by the d descriptor the false alarms appear mainly in homogenous areas, where the variance of the pixel values in the blocks to be compared may be very low, thus the normalized correlation coefficient is highly sensitive to noise. On the other hand, if we consider D and C as a Boolean lattice, where 'true' corresponds to the foreground label, the logical AND operation on D and C improves the results significantly [see Fig. 3.2(h)]. We note that this classification is still quite noisy, although in the segmented image, we expect connected regions representing the motion silhouettes. Applying Markov Random Fields (MRFs) could be a straightforward idea here, however the above introduced label-based 'AND' fusion rule requires a novel three-layer MRF (L^3MRF) model structure that will be introduced in the next section.

3.2.2 Multi-layer segmentation model

In the proposed approach, we construct an MRF model on a graph $\mathcal{G}$ whose structure is shown in Fig. 3.3. In the previous section, we segmented the images in two independent ways, and derived the final result through pixel by pixel label operations using the two segmentations. Therefore, we arrange the sites of $\mathcal{G}$ into three layers S^d, S^c and S^*, each layer has the same size as the image lattice S. We assign to each pixel $s \in S$ a unique site in each layer: e.g. s^d is the site corresponding to pixel s on the layer S^d. We denote $s^c \in S^c$ and $s^* \in S^*$ similarly.

We introduce a labeling process, which assigns a label $\varsigma(.)$ to all sites of $\mathcal{G}$ from the label-set: $L = \{\mathrm{fg}, \mathrm{bg}\}$. The labeling of S^d (resp. S^c) corresponds to the segmentation based on the $f_d(.)$ (resp. $f_c(.)$) feature alone, while the labels at the S^* layer represent the final change mask. A global labeling of $\mathcal{G}$ is $\varpi = \{\varsigma(s^i)|s \in S, i \in \{d, c, *\}\}$.

Following the MRF concept, the labeling of an arbitrary site depends directly on the labels of its neighbors, defined by the neighborhood relations within $\mathcal{G}$. To ensure the smoothness of the segmentations, we put connections within each layer between site pairs corresponding to neighboring pixels of the image lattice S (using 4 neighborhoods). On the other hand, the sites at different layers corresponding to the same pixel must interact in order to produce the fusion of the two different segmentation labels in the S^* layer. Hence, we introduce 'inter-layer' connections between sites s^i and s^j: $\forall s \in S; i, j \in \{d, c, *\}$, $i \neq j$. Therefore, the graph has doubleton 'intra-layer' cliques (their set is $\mathcal{C}_2$) which contain pairs of sites, and 'inter-layer' cliques ($\mathcal{C}_3$) consisting of site-triples. We also use singleton cliques ($\mathcal{C}_1$), which are one-element sets containing the individual sites: they will link the model to the local observations. Hence, the set of cliques is $\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \mathcal{C}_3$.

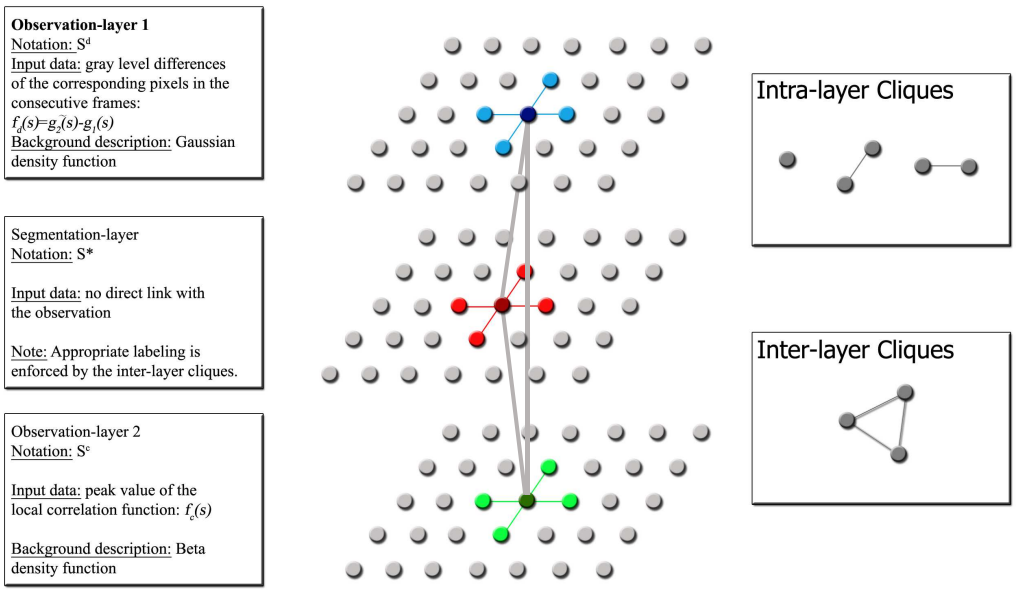


Figure 3.3: Structure of the proposed three-layer MRF (L^3MRF) model

The observation process is defined by $\mathcal{F} = \{\overline{f}(s)|s \in S\}$, where $\overline{f}(s) = [f_d(s), f_c(s)]$. Our goal is to find the optimal labeling $\widehat{\varpi}$, which maximizes the posterior probability: $\widehat{\varpi} = \text{argmax}_{\varpi \in \Upsilon} P(\varpi|\mathcal{F})$, where following the notations from Chapter 2, Υ denotes the set of all possible global labelings. Based on the Hammersley-Clifford Theorem (eq. (2.5)), the a posteriori probability of a given labeling follows a Gibbs distribution.

Our remaining task is to define the V_C *clique potentials*, which have 'low' values if ϖ_C (the label-subconfiguration corresponding to C) is semantically correct, and 'high' otherwise. The observations affect the model through the singleton potentials. As we stated previously, the labels in S^d and S^c layers are directly influenced by the $f_d(.)$ and $f_c(.)$ values, respectively, hence $\forall s \in S$:

$$V_{\{s^d\}}\left(\varsigma(s^d)\right) = -\log P(f_d(s)|\varsigma(s^d)), \qquad V_{\{s^c\}}\left(\varsigma(s^c)\right) = -\log P(f_c(s)|\varsigma(s^c)) \tag{3.1}$$

where the probabilities that the given foreground or background classes generate the $f_d(s)$ or $f_c(s)$ observation have already been defined in Section 3.2.1. Since the labels at S^* have no direct links with the above measurements, uniformly zero potentials can be used there: $V_{\{s^*\}}\left(\varsigma(s^*)\right) = 0$.

In order to get a smooth segmentation at each layer, the potential of an intra-layer clique $C_2 = \{s^i, r^i\} \in \mathcal{C}_2$, $i \in \{d, c, *\}$ favors homogenous labels:

$$V_{C_2} = \Theta\left(\varsigma(s^i), \varsigma(r^i)\right) = \begin{cases} -\delta^i & \text{if } \varsigma(s^i) = \varsigma(r^i) \\ +\delta^i & \text{if } \varsigma(s^i) \neq \varsigma(r^i) \end{cases} \tag{3.2}$$

with a constant $\delta^i > 0$.

As we concluded from the experiments in Section 3.2.1, a pixel is likely to be generated by the background process, if at least one corresponding site has the label 'bg' in the S_d and S_c layers. Its indicator function is noted here by $\mathbf{1}_{\mathrm{bg}} : S^d \cup S^c \cup S^* \to \{0,\ 1\}$, where

$$\mathbf{1}_{\mathrm{bg}}(\upsilon) = \begin{cases} 1 & \text{if } \varsigma(\upsilon) = \mathrm{bg} \\ 0 & \text{if } \varsigma(\upsilon) \neq \mathrm{bg}. \end{cases} \tag{3.3}$$

With this notation and $\rho > 0$ the potential of an inter-layer clique $C_3 = \{s^d, s^c, s^*\}$ is:

$$V_{C_3}(\varpi_{C_3}) = V_{C_3}\left(\varsigma(s^d), \varsigma(s^c), \varsigma(s^*)\right) = \begin{cases} -\rho & \text{if } \mathbf{1}_{\mathrm{bg}}(s^*) = \max\left(\mathbf{1}_{\mathrm{bg}}(s^d), \mathbf{1}_{\mathrm{bg}}(s^c)\right) \\ +\rho & \text{otherwise,} \end{cases} \tag{3.4}$$

Therefore, the optimal MAP labeling $\widehat{\varpi}$, which maximizes $P(\widehat{\varpi}|\mathcal{F})$ (hence minimizes $-\log P(\widehat{\varpi}|\mathcal{F})$) can be calculated using (3.1)−(3.4), and $i \in \{d, c, *\}$ as:

$$\widehat{\varpi} = \underset{\varpi \in \Upsilon}{\operatorname{argmin}} \Big\{ -\sum_{s \in S} \log P(f_d(s)|\varsigma(s^d)) - \sum_{s \in S} \log P(f_c(s)|\varsigma(s^c)) + \\ + \sum_{i;\{s,r\} \in \mathcal{C}_2} \Theta\left(\varsigma(s^i), \varsigma(r^i)\right) + \sum_{s \in S} V_{C_3}\left(\varsigma(s^d), \varsigma(s^c), \varsigma(s^*)\right) \Big\} \tag{3.5}$$

3.2.3 L^3MRF Optimization

The energy term of (3.5) can be optimized by conventional iterative techniques, like ICM [120] or simulated annealing [102]. Accordingly, the three layers of the model are simultaneously optimized, and their interactions develop the final segmentation, which is taken at the end as the labeling of the S^* layer. To obtain a good suboptimal solution, we have developed a three layer modification of the deterministic Modified Metropolis (MMD) algorithm, which provides an efficient trade off between segmentation speed and quality in many applications [107]. The detailed pseudo code of our extended MMD algorithm adapted to the L^{3}MPP segmentation model is given in Appendix B.

3.2.4 Experiments on object motion detection

For quantitative evaluation, we published a new dataset, called the *SZTAKI AirMotion Benchmark*[1], which contains manually generated Ground Truth masks regarding different aerial images. We use three test sets provided by the Hungarian Ministry of Defence Mapping Company©, which contain 83 (=52+22+9) image pairs. Some demonstrating results of the proposed approach are shown in Fig. 3.4. We compared our method to five previous solutions detailed in [12]: *Reddy* [118], which applies a FFT-based similarity alignment; *Farin*'s MRF technique [121], which uses a risk map

[1]Url: `http://mplab.sztaki.hu/remotesensing/airmotion_benchmark.html`

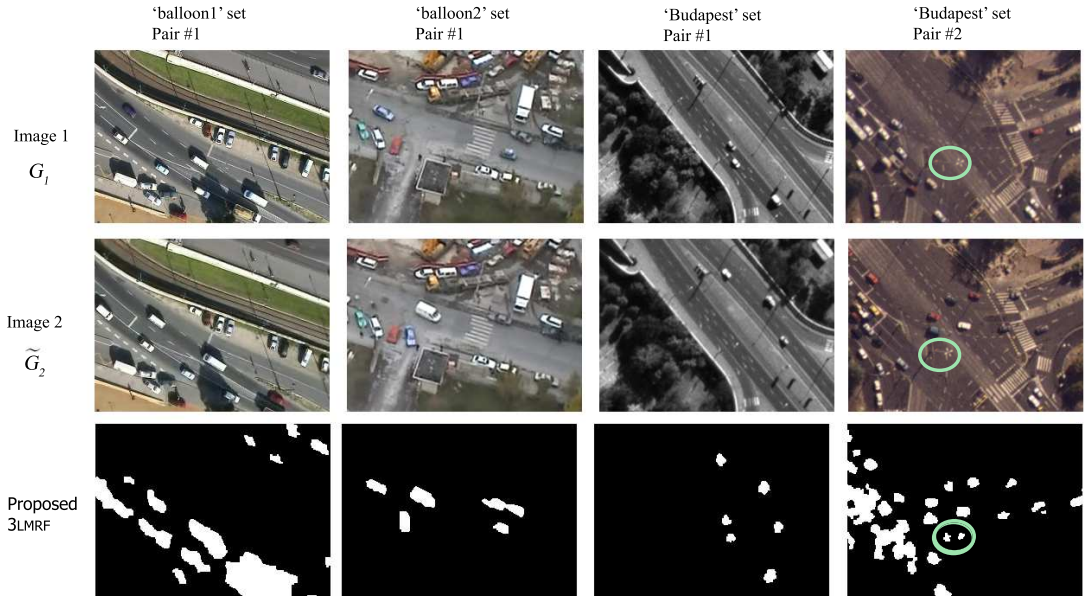


Figure 3.4: Four selected test image pairs for qualitative comparison (see also Fig. B.1)

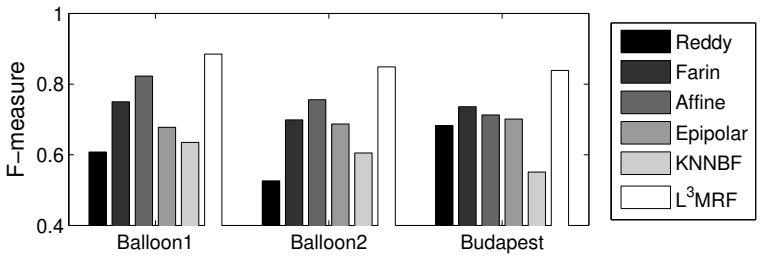
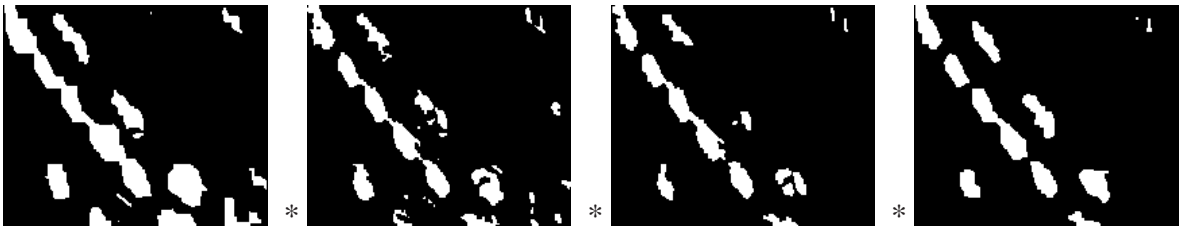


Figure 3.5: Numerical comparison of the proposed model (L^3MRF) to five reference methods, using three test sets: 'balloon1' (52 image pairs), 'balloon2' (22) and 'Budapest' (9).

to decrease the registration errors; a keypoint based *Affine* transform estimator [122, 123]; the *Epipolar* approach [124], where each pixel is checked against the homography and epipolar [125] constraints, labeling outliers of both comparisons as foreground; and the *K-Nearest-Neighbor-Based Fusion Procedure (*KNNBF*)* motion segmentation method [126] which is one of the main applications of the label fusion framework [138]. In the quantitative experiments, we investigated on how many pixels have the same label in the Ground Truth masks and in the segmented images obtained by the different methods. For evaluation criterion, we use the *F-score* which combines *Recall* and *Precision* of foreground detection in a single efficiency measure. Results in Fig. 3.5 show the superiority of the proposed L^3MRF model versus previous approaches.

Another relevant issue of validation is to compare the proposed L^3MRF model structure - in the context of the addressed application - to different information fusion approaches. As demon-

(a) Image 1 (G_1) (b) Image 2 (G_2) (c) GT motion regions (d) Observation fusion



(e) Decision fusion (f) L^3MRF-δ_0^* (g) L^3MRF-$\delta_0^{d,c}$ (h) Proposed L^3MRF

Figure 3.6: Evaluation of the proposed L^3MRF model versus different fusion approaches.

strated in Fig. 3.6, our experiments confirmed the benefits of the introduced L^3MRF structure for the addressed problem versus four different information fusion models. It has been shown that the 2D joint density representation of the two examined features (called *Observation fusion*) cannot appropriately express here the desired relationship between the feature and label maps, while the multi-layer approach provides an efficient solution. On the other hand, considering the task as a global Bayesian optimization problem (3.5) is preferred versus applying a sequential *Decision fusion* process, where two independent label maps are created first based on the $f_d(.)$ and $f_c(.)$ features respectively, thereafter, the segmentation of S^* is derived by a *pixel by pixel* AND operation from the two change maps. Finally, using intra-layer smoothing interactions in each layer contributes to the improved segmentation result, unlike in the two following variants: L^3MRF-δ_0^*, which uses $\delta^* = 0$ settings, and L^3MRF-$\delta_0^{d,c}$ using $\delta^d = 0$ and $\delta^c = 0$.

3.3 Long term change detection in aerial photos

This section focuses on change detection in optical aerial images which were taken with several years time differences partially in different seasons and lighting conditions (Fig. 3.7). In this case, straightforward techniques like thresholding the difference image [12, 127] cannot be efficiently adopted since the observed pixel levels even in the 'unchanged' image regions may be significantly different, due to different illumination conditions, or seasonal changes in vegetation.

In our approach, similarly to Sec. 3.2, changes are identified through two complementary descriptors, however the label fusion part will be more complex than in L^{3}MRF. Instead of combining the two feature-based label maps via logical operators, we utilize a third feature, which is respon-

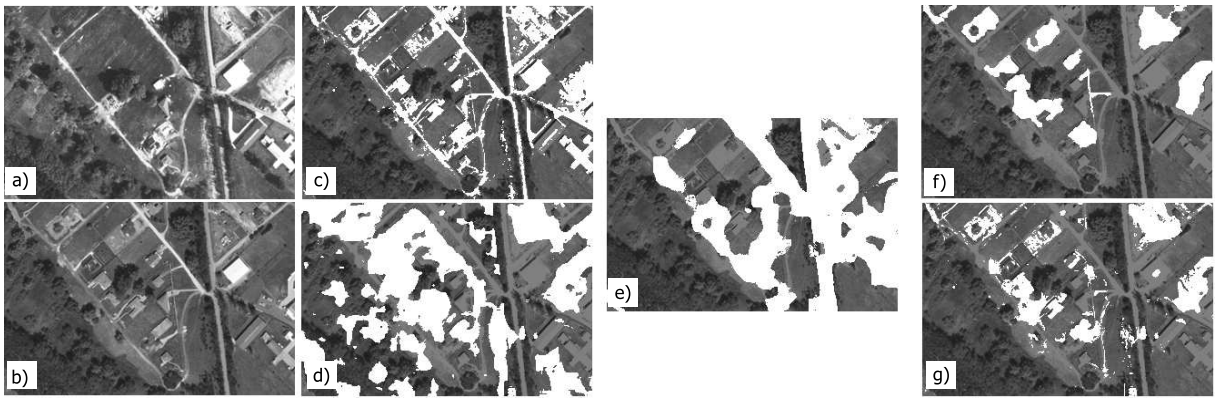


Figure 3.7: Feature selection for long term change detection: a) image 1 (G_1), b) image 2 (G_2), c) intensity based change detection ($\phi_g(.)$, changes are marked with white), d) correlation based change detection ($\phi_c(.)$), e) local variance based segmentation, white if $\phi_\nu(s) = c$, f) Ground Truth, g) change detection results obtained by per pixel integration of $\phi_g(.)$, $\phi_c(.)$ and $\phi_\nu(.)$ maps

sible for locally choosing the more reliable change descriptor in the different image regions. This modification requires involving dynamic connections between the nodes of the multi-layer graph structure, that will be implemented using the Mixed Markov model concept [45, 75].

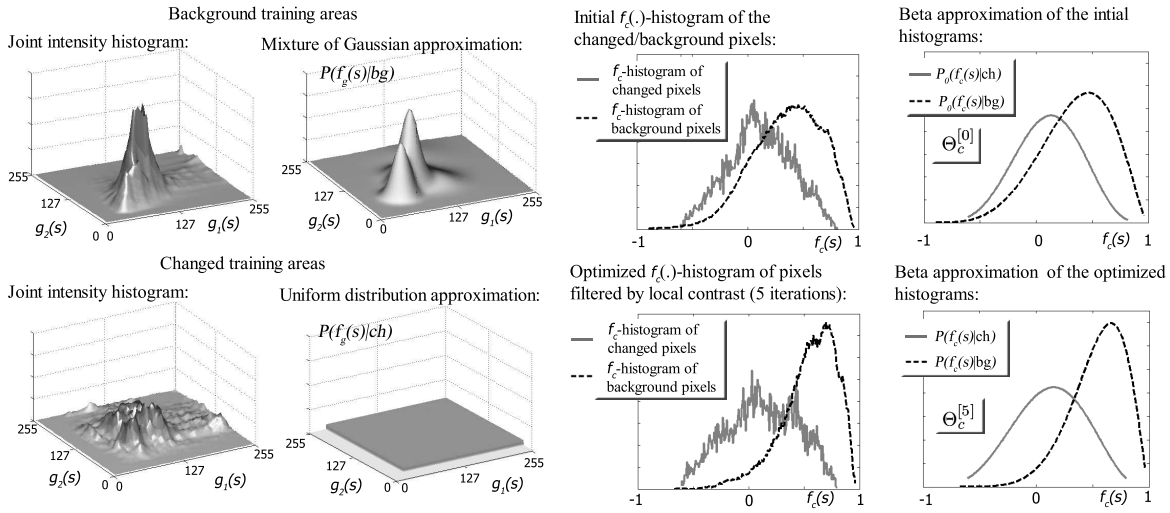
For simplicity, we use here various notations from Sec. 3.2. Let G_1 and G_2 be again the two input (grayscale) images, but we assume now that G_1 and G_2 have an identical pixel lattice S and they are already registered by the image providers. The later assumption is reasonable since in contrast to object motion detection, long term change detection is an offline task. The gray values are henceforward denoted by $g_1(s)$ and $g_2(s)$ for a pixel $s \in S$ of G_1 and G_2, respectively.

3.3.1 Image model and feature extraction

We start our investigations in the joint intensity domain of the two images. Let us consider the 2D histogram of the $f_g(s) = [g_1(s), g_2(s)]^T$ vectors extracted over the *background* regions of the training images [see Fig 3.8(a) regarding the image pair of Fig. 3.7]. Thereafter we approximate this histogram by a mixture of K Gaussian distributions, where K is a parameter of the model. In this way, we measure which intensity values occur often together in the corresponding images. Thus the probability of the $f_g(s)$ observation in the background is calculated as:

$$P\big(f_g(s)\big|\mathrm{bg}\big) = \sum_{i=1}^{K} \kappa_i \cdot \eta\Big(f_g(s), \overline{\mu}_i, \overline{\overline{\Sigma}}_i\Big),$$

where $\eta(.)$ denotes a two dimensional Gaussian density function with $\overline{\mu}_i$ mean vector and $\overline{\overline{\Sigma}}_i$ covariance matrix, while the κ_i terms are positive weighting factors ($\sum_{i=1}^{K} \kappa_i = 1$). Fig 3.8(a) shows the Expectation Maximization (EM) estimate [128] of the density using $K = 5$ mixture



(a) f_g feature histograms (b) f_c feature histograms

Figure 3.8: Feature histograms with statistical approximations

components. While the background's intensity model exploits the presence of a few frequently co-occuring gray level pairs in the two images (e.g. the mean color of plough lands or forests), the $f_g(s)$ histogram of the changed regions has usually several smaller peaks covering a significant part of the 2D intensity domain. Expressing that any $f_g(s)$ value may occur in the changed areas with similar probabilities, the 'ch' class is modeled by a uniform density $P\big(f_g(s)\big|\mathrm{ch}\big)$ [129].

Obviously, the $f_g(s)$ feature cannot separate alone changed and unchanged regions. As Fig. 3.7(c) shows, the obtained $f_g(s)$-based maximum likelihood label map $\phi_g : S \to \{\mathrm{ch}, \mathrm{bg}\}$ contains several false changes in unaltered territories, mainly in highly textured regions (e.g. areas of buildings and roads), where the occurring $f_g(s)$ gray value pairs are less frequent in the global image statistics.

Similarly to the object motion detection application in Sec. 3.2, the second feature $c(s)$ is calculated as the normalized cross correlation between the $z \times z$ neighborhoods of pixel s in G_1 and G_2 images, respectively. In Fig 3.8(b), we plot the histogram of the obtained $f_c(s)$ values over the changed respectively background regions of the training images. Considering the asymmetry of the empirical distributions, we have found that Beta density approximations [119] are appropriate for the classes: $P(f_c(s)|\mathrm{ch}) = B\big([f_c(s)+1]/2, \alpha_{\mathrm{ch}}, \beta_{\mathrm{ch}}\big)$ and $P\big(f_c(s)\big|\mathrm{bg}\big) = B\big([f_c(s)+1]/2, \alpha_{\mathrm{bg}}, \beta_{\mathrm{bg}}\big)$.

As the $c(s)$-based $\phi_c : S \to \{\mathrm{ch}, \mathrm{bg}\}$ maximum likelihood segmentation result in Fig. 3.7(d) shows, this feature is also weak in itself. However we should observe that $f_g(s)$ and $f_c(s)$ are efficient complementary features. In low contrasted, homogeneous image regions, where the noisy $f_c(s)$ may be irrelevant, the decision based on $f_g(s)$ seems to be fairly reliable. On the other hand

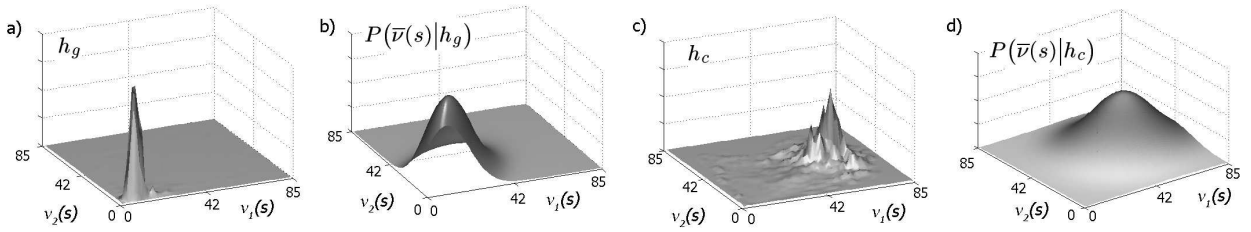


Figure 3.9: Illustration of the 2 dimensional h_g and h_c histograms as function of the corresponding $\nu_1(s)$ and $\nu_2(s)$ values

in textured areas one should choose $f_c(s)$ instead of $f_g(s)$.

In the following, we formulate the *contrast based feature selection* in a probabilistic manner. We measure the local contrast over image G_i by $\nu_i(s)$ ($i \in \{1,2\}$), as the variance of the gray levels in a rectangular neighborhood of s. Let be $\overline{\nu}(s) = [\nu_1(s), \nu_2(s)]^T$. We denote by T the Ground Truth mask with $t(s) \in \{\mathrm{ch}, \mathrm{bg}\}$ labels $\forall s \in S$, and δ is the Kronecker-delta.

Next, we quantitatively examine the correspondence between the observed $\overline{\nu}(s)$ value and the ML classification performance using the $f_g(s)$ and $f_c(s)$ features, respectively. We particionate the domain of the occurring $\nu_1(s)$ [similarly $\nu_2(s)$] values with L equal bins: $b_1, \ldots, b_L$ (each b_n is a line segment in $\mathbb{R}$.) We say that $\overline{\nu}(s) \in \overline{b}_{m,n}$ if $\nu_1(s) \in b_m$ and $\nu_2(s) \in b_n$ ($\overline{b}_{m,n}$ is a rectangle in $\mathbb{R}^2$). Next we build the following *ratio histogram* h_g, which measures for each $\overline{b}_{m,n}$ bin the ratio of the number of correctly and erroneously classified pixels through $\phi_g(.)$, where the corresponding $\overline{\nu}(s)$ values lie in $\overline{b}_{m,n}$. With $S_{m,n} = \{s | s \in S, \overline{\nu}(s) \in \overline{b}_{m,n}\}$:

$$h_g[m,n] = \frac{\sum_{s \in S_{m,n}} \delta\big(t(s), \phi_g(s)\big)}{\sum_{s \in S_{m,n}} \Big(1 - \delta\big(t(s), \phi_g(s)\big)\Big)} \tag{3.6}$$

h_c can be defined similarly for the $f_c(.)$ feature.

We illustrate the h_g and h_c 2D *ratio histograms* in Fig. 3.9(a) and 3.9(c). High peaks of h_g (resp. h_c) indicate domains of $\overline{\nu}(s)$ where the decision based on the $f_g(.)$ [resp. $f_c(.)$] feature is reliable. After normalization, the histograms can be considered as probability distributions which we approximate again with parametric density functions. In this case, the two classes being modeled are g and c, indicating the $\overline{\nu}(s)$ domains where the $f_g(s)$ respectively $f_c(s)$ features are more reliable regarding the ch/bg classification of pixel s. In the experiments the two domains proved to be fairly separable with 2D Gaussian density approximations of the h_g and h_c histograms as it is shown in Fig. 3.9(b) and 3.9(d) (the histograms are unimodal and only slightly overlapping). Thus we use the following distributions: $P\big(\overline{\nu}(s)\big|h_g\big) = \eta\Big(\overline{\nu}(s), \overline{\mu}_g, \overline{\overline{\Sigma}}_g\Big)$ and $P\big(\overline{\nu}(s)\big|h_c\big) = \eta\Big(\overline{\nu}(s), \overline{\mu}_c, \overline{\overline{\Sigma}}_c\Big)$. Thereafter we can obtain the ML contrast map [Fig. 3.7(e)]

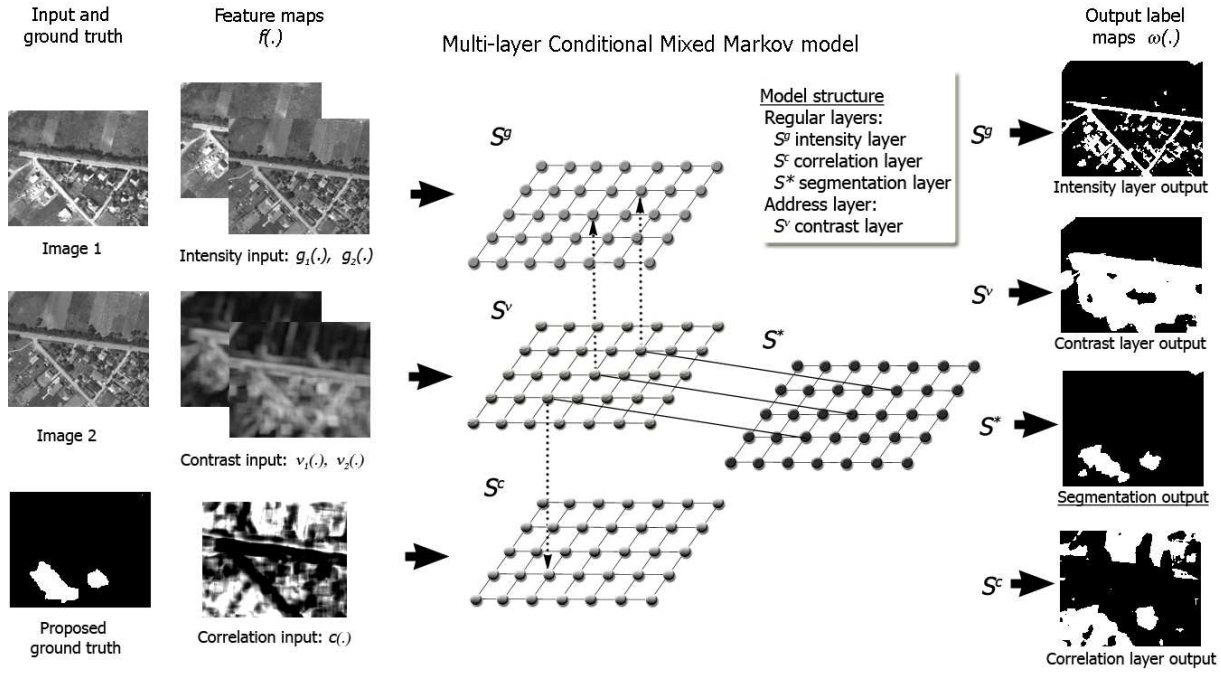


Figure 3.10: Structure of the proposed model and overview of the segmentation process.

as: $\phi_\nu(s) = \text{argmax}_{\chi\in\{g,c\}} P\big(\overline{\nu}(s)\big|h_\chi\big)$. For estimating the final change mask, ϕ_*, the following pixel-by-pixel segmentation process can be taken:

$$\phi_*(s) = \begin{cases} \phi_g(s) & \text{if } \phi_\nu(s) = g \\ \phi_c(s) & \text{if } \phi_\nu(s) = c \end{cases} \tag{3.7}$$

As Fig. 3.7(g) shows, pixel-by-pixel segmentation based on eq. (3.7) is quite noisy, calling for Markovian filtering of the change mask. However incorporating the above feature-selection-based segmentation schema needs a different approach from the earlier introduced L^3MRF model, as the $\overline{\nu}(s)$ feature plays a particular role: it can locally switch ON and OFF the $f_g(s)$ respectively $f_c(s)$ features into the integration process. Since in MRFs the interactions between the processing nodes must be static, we will implement an extended structure using the mixed Markov model concept [75], which will be investigated in the next section.

3.3.2 A Conditional Mixed Markov image segmentation model

The proposed approach, called the Conditional MiXed Markov Model (CXM), is a combination of a mixed Markov model [75] and a conditionally independent random field of signals. We map the problem onto a graph $\mathcal{G}$ whose structure is shown in Fig. 3.10(c). Previously, we segmented the images in three different ways, and derived the final result through pixel-by-pixel label operations

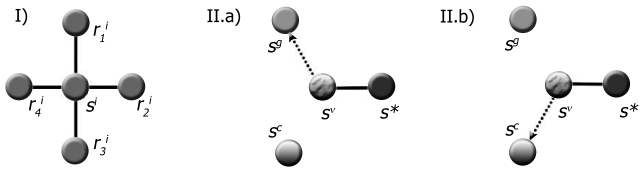


Figure 3.11: Demonstration of (I) intra- and (II.a,II.b) inter-layer connections regarding nodes associated to pixel s. Continuous line is an edge of $\mathcal{G}$, dotted arrows denote the two possible a destinations of the address node s^ν. (in I: $i \in \{g, c, \nu, *\}$)

using the three segmentations. Therefore we arrange the nodes of $\mathcal{G}$ into four layers: S^g, S^c, S^ν and S^*, where each layer has the same size as the S image lattice. S^g, S^c and S^ν are called the *feature layers*, and S^* is the combined segmentation layer. We assign to each pixel $s \in S$ a unique node in each layer: e.g. s^g is the node corresponding to pixel s on the layer S^g. We denote $s^c \in S^c$, $s^\nu \in S^\nu$ and $s^* \in S^*$ similarly.

First step is the definition of the labeling random process, which assigns a label $\varsigma(q)$ to each q node of $\mathcal{G}$. As usual in mixed models [75], graph edges and address pointers express direct dependencies between the corresponding node labels. The S^g, S^c, and S^* layers of the model contain *regular nodes*, where the label denotes a possible ch/bg segmentation class:

$$\forall s \in S, i \in \{g, c, *\} : \ \varsigma(s^i) \in \{\mathrm{ch}, \mathrm{bg}\} \tag{3.8}$$

For each s, $\varsigma(s^g)$ resp. $\varsigma(s^c)$ corresponds to the segmentation directly influenced by the $f_g(s)$ resp. $f_c(s)$ feature; while the labels at the S^* layer present the final change mask.

On the other hand the S^ν layer is responsible for matching the regions of the final change map S^* to appropriately segmented regions either in the S^g or in the S^c layers. Hence S^ν will be an *address layer*, with node-pointer labels $\{\varsigma(s^\nu) | \forall s \in S\}$.

Next we describe how the model encapsulates the information extracted from the input images. We use a $f(.)$ operator which assigns to the nodes of the *feature layers* S^g, S^c and S^ν the corresponding local observations, so that $f(s^g) = f_g(s)$, $f(s^c) = f_c(s)$ and $f(s^\nu) = \overline{\nu}(s)$, $\forall s \in S$. Denote the global observation process by $\mathcal{F} = \{f(q) | q \in \mathcal{O}\}$, where $\mathcal{O} = S^g \cup S^c \cup S^\nu$.

Similarly to MRFs, our proposed CXM segmentation model follows the Maximum a Posteriori (MAP) approach [12, 102], looking for the global labeling $\widehat{\varpi}$ which maximizes the following conditional probability: $P(\varpi|\mathcal{F}) = P(\mathcal{F}|\varpi) \cdot P(\varpi)$. Assuming conditionally independent observations, $P(\mathcal{F}|\varpi)$ can be obtained as a product of $P\big(f(q)|\varsigma(q)\big)$ *singleton* probability terms assigned to the nodes of the feature layers. In the S^g and S^c layers, we calculate the node-by-node singletons using the same probability density functions which have already been defined in Section 3.3.1. Thus $\forall s \in S$ and $\psi \in \{\mathrm{ch}, \mathrm{bg}\}$:

$$P\big(f(s^g)|\varsigma(s^g) = \psi\big) = P\big(f_g(s)|\psi\big) \quad P\big(f(s^c)|\varsigma(s^c) = \psi\big) = P\big(f_c(s)|\psi\big) \tag{3.9}$$

Singletons of S^ν will be defined later.

On the other hand using CXM the $P(\varpi)$ prior probability derives from a mixed Markov model, thus it follows eq. (2.14). To calculate $P(\varpi)$, we have to define appropriately the edges (or cliques) of $\mathcal{G}$ and the corresponding V_C clique potential functions. To fulfill the desired constraints, we use in the model two types of cliques representing intra- and inter-layer interactions (see Fig. 3.11).

For the sake of obtaining smooth segmentations, we put connections within each layer among node pairs corresponding to (4-)neighboring pixels on the S image lattice. Denote the set of the resulting *intra-layer* cliques by $\mathcal{C}_2$. The prescribed potential function of a clique in $\mathcal{C}_2$ penalizes neighboring nodes having different labels. Assuming r and s to be neighboring pixels on S, the potential of the doubleton clique $C_2 = \{r^i, s^i\} \in \mathcal{C}_2$ for each $i \in \{g, c, \nu, *\}$ is calculated as:

$$V_{C_2}\Big(\varsigma(s^i), \varsigma(r^i)\Big) = \begin{cases} -\delta^i & \text{if } \varsigma(s^i) = \varsigma(r^i) \\ +\delta^i & \text{if } \varsigma(s^i) \neq \varsigma(r^i) \end{cases} \tag{3.10}$$

with a constant $\varphi^i > 0$.

Now let us continue with the description of the inter-layer interactions. Based on previous investigations [see (3.7)], $\varsigma(s^*)$ should mostly be equal either to $\varsigma(s^g)$ or to $\varsigma(s^c)$, depending on the 'vote' of the $\nu(s)$ feature. Hence we put an edge among s^* and s^ν as well as we prescribe that address node s^ν should point either to s^g or to s^c:

$$\forall s \in S: \ \varsigma(s^\nu) \in \{s^g, s^c\} \tag{3.11}$$

The directions of the address pointers are influenced by the singletons of S^ν where we use the distributions defined in Sec. 3.3.1:

$$P\big(f(s^\nu)|\varsigma(s^\nu) = s^\chi\big) = P\big(\overline{\nu}(s)|h_\chi\big), \ \ \chi \in \{g, c\} \tag{3.12}$$

Finally we get the potential function of the inter-layer clique $C_3 = \{s^*, s^\nu\}$ as

$$V_{C_3}\Big(\varsigma(s^*), \tilde{\varsigma}(s^\nu)\Big) = \begin{cases} -\rho & \text{if } \varsigma(s^*) = \tilde{\varsigma}(s^\nu) \\ +\rho & \text{otherwise} \end{cases} \tag{3.13}$$

where $\rho > 0$, and using (2.13): $\tilde{\varsigma}(s^\nu) = \varsigma\big(\varsigma(s^\nu)\big)$.

Using the above introduced energy terms the optimal $\widehat{\varpi}$ can be calculated as:

$$\widehat{\varpi} = \underset{\varpi \in \Omega}{\operatorname{argmin}} \Bigg\{ \sum_{s \in S} -\log P\big(f_g(s)|\varsigma(s^g)\big) + \sum_{s \in S} -\log P\big(f_c(s)|\varsigma(s^c)\big) + \sum_{s \in S} -\log P\big(\overline{\nu}(s)|\varsigma(s^\nu)\big) + \sum_{i;\{s,r\} \in \mathcal{C}_2} V_{C_2}\big(\varsigma(s^i), \varsigma(r^i)\big) + \sum_{s \in S} V_{C_3}\big(\varsigma(s^*), \tilde{\varsigma}(s^\nu)\big) \Bigg\} \tag{3.14}$$

where $i \in \{g, c, \nu, *\}$ and Ω denotes the set of all the possible global labelings.

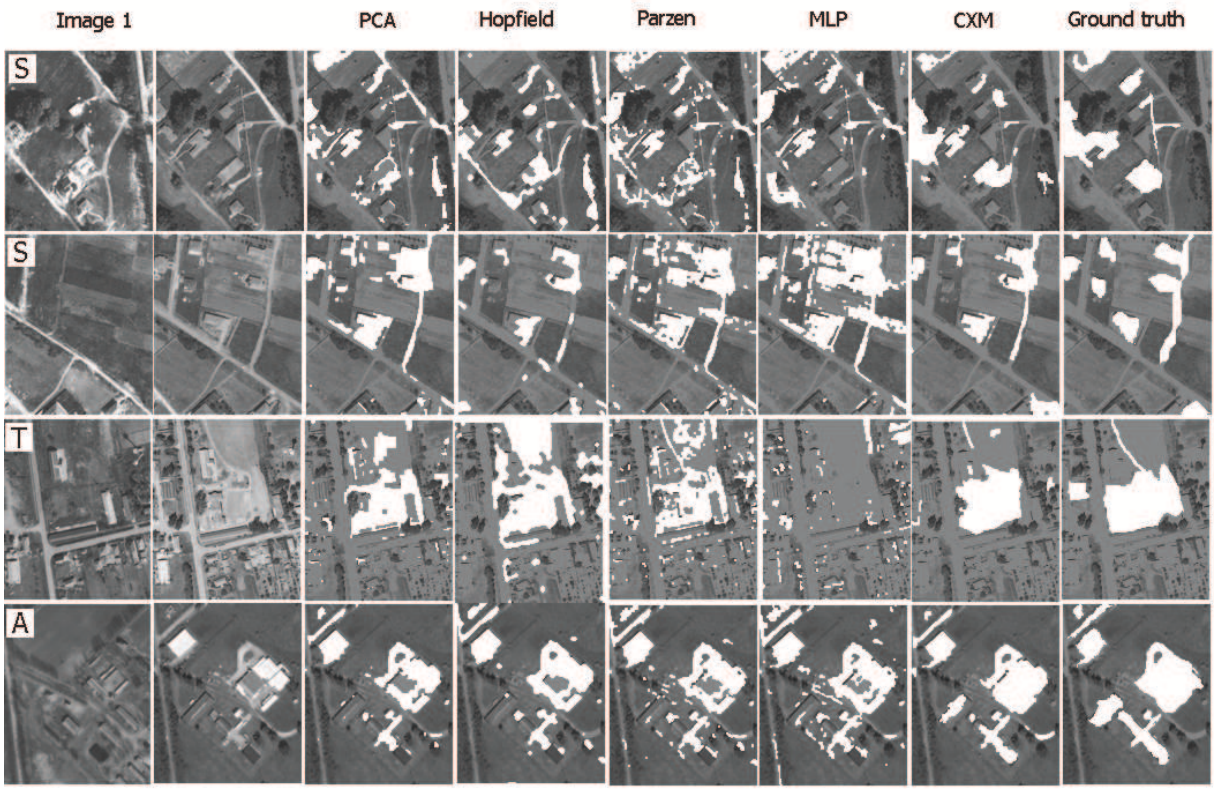


Figure 3.12: Qualitative comparison of the change detection results with the different test methods and the proposed CXM model, for data sets: SZADA (S), TISZADOB (T) and ARCHIVE (A). White regions mark the detected/Ground Truth changes.

We minimize the energy term of eq. (3.14) again with the deterministic Modified Metropolis relaxation process, in a similar manner to the L^3MRF's optimization algorithm presented in Appendix B. Note that due to its fully modular structure, the introduced model could be completed in a straightforward way with additional sensor information (e.g. color or infrared sensors) or task-specific features depending on availability.

3.3.3 Experiments on long term change detection

For evaluation of CXM we used *three* sets of optical aerial image pairs provided by the Hungarian Institute of Geodesy Cartography & Remote Sensing (FÖMI) and Google Earth (see Fig. 3.12). We published the labeled test data as the *SZTAKI AirChange Benchmark Set*[1].

Data set SZADA contains images taken by FÖMI in 2000 and in 2005, respectively. This test set consists of *seven* - also manually evaluated - photo pairs, covering in aggregate 9.5km^2 area at *1.5m/pixel* resolution (the size of each image in the test set is 952×640 pixels). One image

[1]Url: `http://mplab.sztaki.hu/remotesensing/airchange_benchmark.html`

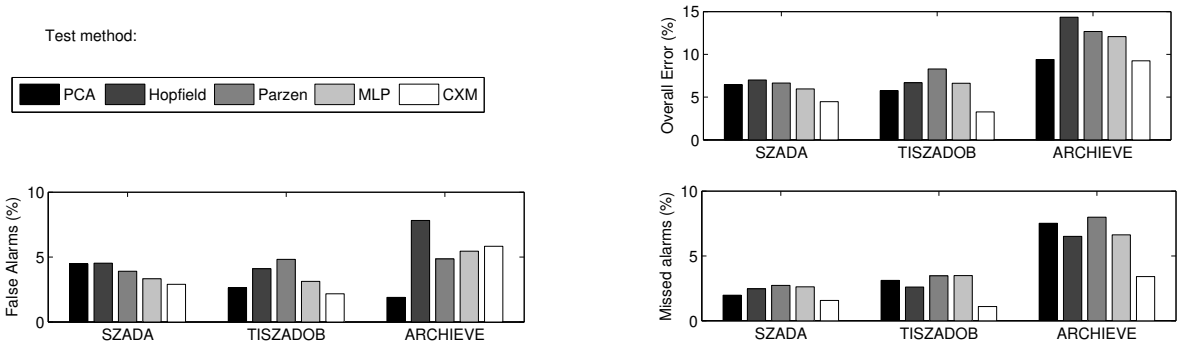


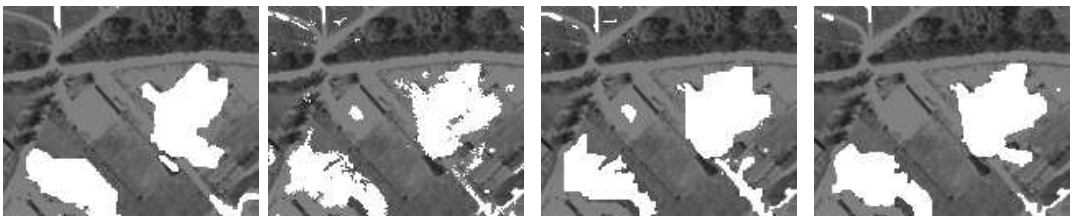
Figure 3.13: Quantitative comparison of the proposed CXM technique to four previous methods on the three sets of the *SZTAKI AirChange Benchmark*: SZADA, TISZADOB and ARCHIVE. False alarm, missed alarm and overall error rates are given in percent of the checked pixels.

pair has been used here for training and the remaining six ones for validation. The second test set called TISZADOB includes *five* photo pairs from 2000 resp. 2007 ($6.8\mathrm{km}^2$) with similar size and quality parameters to SZADA. Finally, in the test ARCHIVE, we compared an aerial image taken by FÖMI in 1984 to a corresponding Google Earth photo from around 2007. The latter case is highly challenging, since the photo from 1984 has a poor quality, and several major differences appear due to the 23 years time difference between the two shots. Ground Truth masks have been generated manually for each image pair of the training and test sets. The following changes have been considered: new built-up regions, building operations, planting forests or individual trees (trees only at high resolution), fresh plough-lands and groundworks before building over.

In our paper introducing CXM [13], we quantitatively compared our method to four previous solutions: the first technique, called *PCA* [130], projects the joint gray level vectors into the space of the principal components estimated over the background training regions, and applies MRF classifications. The remaining three references, *Hopfield* [131], *Parzen* [132] and *MLP* [133, 134] segment the difference image with different supervised techniques. During the numerical tests, we used the same metric as [131, 132]: we compared the segmentation results provided by the different techniques to the Ground Truth (GT), and measured the numbers of false alarms (unchanged pixels which were detected as changes), missed alarms (erroneously ignored changed pixels) and overall error (sum of the previous two quantities). Comparative results on the three test sets are given in Fig. 3.13 in percent of the number of processed image pixels. As this figure shows, the overall error of the proposed CXM model was below the error of the reference methods by about $2-5$ percent. Note that the generally weaker results in the ARCHIVE tests were primarily caused by the lower image quality. For the sake of visual demonstration, we show the comparative change detection results of some relevant image parts in Fig. 3.12. We can observe that the CXM model produced smooth and more accurate change regions in the selected areas, compared to the reference methods.

In Fig. 3.14, similarly to the evaluation of L^3MRF, we also demonstrate that the introduced label-based feature integration approach outperformed decision fusion techniques.

Since the publication date of the CXM model in 2009 [13], various novel multi-layer MRF models have been published. For this reason, in 2015 we prepared an up-to-date survey article on the existing techniques [4], comparing CXM in details to two newer approaches [117, 135] from 2014, in cooperation with their authors. Here the first reference is a *Multicue MRF* model [135], which integrates the modified Histogram of Oriented Gradients and graylevel difference features into the original Multi-MRF structure framework proposed by [116], where two layers correspond to the two feature maps and the third one is the final segmentation layer. The class models and the inter-layer interaction terms are both affected by observation dependent and prior constraints, differently from CXM where the feature maps only affect the singleton terms, while the interaction terms implement purely prior label fusion (soft-)constraints. The *second* reference, called *Fusion-MRF* [117], simultaneously realizes adaptive segmentation and change detection for optical remote sensing images, where each layer represents a given input image; thus "multi-layer" refers here to multi-temporal images. In our study [4], the quantitative comparison was based again on the *SZTAKI AirChange Benchmark Set*, however depending on the different approaches on change modeling, further considerations should have been taken. *Multicue MRF* is a *direct* change detection technique similarly to our CXM [13] and all reference methods of Fig. 3.13, which obtain changes through segmenting similarity maps between the input images. On the other hand the *Fusion MRF* follows a Post Classification Comparison (PCC) approach, which segment first the input images into various land-cover classes, and changes are obtained indirectly as regions with different class labels in the different time layers. Therefore, by testing the *Fusion MRF*, we could not rely on the available Ground Truth (GT) change masks (referred as *AirChange GT*). Instead, we generated new GT (called *Region PCC GT*), where various land-cover classes have been considered for different image pairs of the test set, such as urban and non-urban; or meadow, planted meadow and forest. Note that as concluded in [4] the two GT generating approaches correspond to two different use-cases, and both ones may be relevant. Comparative experiments between *CXM* and the *Fusion MRF* with the two different GT types gave obviously different results, showing the superiority of *CXM* with 4-58% in *F-score* using the original *AirChange GT* masks, and the advantages of the *Fusion MRF* (with 14-31%) with the *Region PCC GT*. Calculating the *Overall error* for five selected image pairs (Fig. 3.13) showed the minor superiority of the newer *Multicue MRF* (with a margin of 0.5-2%), while calculating the traditional *Precision*, *Recall* and *F-score* values indicated notable advantages of CXM due to a significantly higher Recall rate. In summary, the experiments of our survey demonstrated that the proposed CXM technique proved to be also competitive versus more recent multi-layer change detection models.



(a) Ground Truth (b) Decision fusion (c) Ensemble of MRFs (d) Proposed CXM

Figure 3.14: Impacts of the multi-layer CXM structure for the quality of the change mask. We compare the results of (b) the pixel-by-pixel classification without spatial smoothing, (c) the ensemble of three independent, single-layer MRFs and (d) the proposed multi-layer model

3.4 Parameter settings in multi-layer segmentation models

The parameters of the introduced multi-layer segmentation models can be divided into three groups: (i) preliminary parameters of feature calculation, (ii) parameters of the probability density functions in the data terms and (iii) parameters of the prior intra- and inter-layer potential functions.

First, the size of block matching windows (z) used for correlation calculation and in L^3MRF the size of search window (l) are related to a priori knowledge about the image resolution and textureness, object size and magnitude of the parallax distortion. The correlation window should not be significantly larger than the expected objects to ensure low correlation between an image part which contains an object and one from the same 'empty' area.

Second, the feature distribution (pdf) parameters can be obtained by conventional Maximum Likelihood estimation algorithms from background and foreground training areas. If manually labeled training data is not available, the foreground training regions must be extracted through outlier detection [136] in the feature spaces.

Third, while data-based pdf parameters strongly depend on the input image data, interaction potential factors are largely independent of it. Experimental evidence suggests that the model is not sensitive to a particular setting of ρ or δ^i within a wide range, which can be estimated a priori. The parameters of the intra-layer potential functions, δ^i, influence the size of the connected blobs in the segmented images. Although automatic estimation methods exist for similar smoothing terms [137], δ^i is rather a hyper-parameter, which can be fixed by trial and error. Higher δ^i values result in more compact foreground regions, however, fine details of the silhouettes may be distorted that way. We have used $\rho = \delta^*$: this choice gives the same importance to the intra-layer smoothness and the inter-layer label fusion constraints.

3.5 Conclusions of the chapter

In this chapter, we have introduced novel Markovian label fusion models for two different change detection problems from the remote sensing domain. *First*, we proposed a three-layer MRF model (L^3MRF) for extracting the regions of object motions from image pairs taken by an airborne moving platform. The output of the proposed method is a *change map*, which can be used e.g. for estimating the dominant motion tracks (i.e. roads) in traffic monitoring tasks, or for outlier region detection in mosaicking and in stereo reconstruction. Moreover, it can also provide an efficient preliminary mask for higher level object detectors and trackers in aerial surveillance applications.

We have shown that even if the preliminary image registration was relatively coarse, the false motion alarms could be fairly eliminated with the integration of frame-differencing with local cross-correlation, which presented complementary features for detecting static scene regions. The efficiency of the method has been validated through three different sets of real-world aerial images, and its behavior versus five reference methods and four different information fusion models has been quantitatively and qualitatively evaluated. The experiments have shown that the proposed model outperformed the reference methods dealing with image pairs with large camera and object motions and significant but bounded parallax.

In the *second* part of the chapter, we addressed the detection of statistically unusual changes in optical aerial image pairs taken with significant time differences. A novel Conditional Mixed Markov (CXM) model has been proposed, which could integrate the robustness of MRF-based segmentation techniques [102], the modularity of multi-layer approaches [138] and semantic flexibility of mixed Markov models [75]. The introduced method utilized information from three different observations: global intensity statistics, local correlation and contrast. The performance of the method has been validated using real-world aerial images. The superiority of CXM versus four earlier reference methods, and its relevance against two newer methods has been shown quantitatively and qualitatively.

Both models of this chapter may be used as efficient and scalable change detection filters for several remote sensing applications. The methods are purely based on low-level features, working without object extraction or identification of land cover classes. Therefore they can be adopted for a large variety of scenes and purposes, even in situations where the concept of ‘interesting changes’ is not well defined. The methods can support manual evaluation of large data sets by focusing the operator’s attention to targets or changed areas, and also in automated systems with restricting the field of interest and presenting shape or region based descriptors for higher level image interpretation modules.

Chapter 4

Multitemporal data analysis with Marked Point Processes

In this chapter we introduce new approaches for object level dynamic scene modeling based on multitemporal measurements, by extending the conventional Marked Point Process framework with modules focusing on the time dimension. First, a new probabilistic method is proposed which integrates building extraction with change detection in pairs of remotely sensed images captured with several years time differences. The output of the method is a population of 2D building footprint segments, where status information is provided for each segment highlighting changes between the two time layers.

In the second part, we propose a Multiframe Marked Point Process model of line segments and point groups for automatic target structure extraction and tracking in Inverse Synthetic Aperture Radar (ISAR) image sequences. For the purpose of dealing with scatterer scintillations and high speckle noise in the ISAR frames, we obtain the resulting target sequence by an iterative optimization process, which simultaneously considers the observed image data and various prior geometric interaction constraints between the target appearances in the consecutive frames.

For both models, detailed quantitative evaluation is performed on real remotely sensed measurements.

4.1 Introducing the time dimension in MPP models

Conventional Marked Point Process (MPP) techniques are applicable for the analysis of static scenarios, however several applications require object level investigations on multitemporal measurements. Our key contribution in this chapter is to propose methodologies for incorporating the time dimension into the MPP framework. We will address two different challenges: object level change detection and moving target analysis. Although both issues are quite general, for easier discussion and validation, we introduce the new model structures for selected applications: building development monitoring and moving object analysis in radar (ISAR) image sequences.

4.2 Object level change detection

In this section we introduce a novel Multitemporal Marked Point Process (mMPP) model, which is able to detect objects and mark the object level changes in remotely sensed image pairs taken at different time instances. We present methodological contributions in three key issues:

- We implement a novel object-change modeling approach, which simultaneously exploits low level change information between the time layers and object level description to recognize and separate changed and unaltered objects.
- Answering the challenges of *data heterogeneity* in aerial and satellite image repositories, we construct a flexible hierarchical framework which can create various object appearance models from different elementary feature based modules.
- To simultaneously ensure the convergence, optimality and computation complexity constraints raised by the increased *data quantity* in remote sensing applications, we adopt the quick *Multiple Birth and Death* optimization technique for change detection purposes, and propose a novel non-uniform stochastic object birth process, which generates relevant objects with higher probability based on low-level image features.

4.2.1 Building development monitoring - problem definition

Following the evolution of built-up regions is a key issue of aerial and satellite image analysis. Although the topic has been extensively studied since the 80's, it has had to continuously face the challenges of the quickly evolving quality and quantity of remotely sensed data, the richness of different building appearances, the data-heterogeneity in the available image repositories and the various requirements of new application areas [40]. As discussed in Chapter 3 (Sec. 3.3), pixel level change detection approaches, such as our conditional mixed Markov model (CXM) [13] can

(a) Input image 1 (b) Input image 2 (c) Changed regions of image 2

Figure 4.1: Low level change detection: (a) and (b) input images, (c) change mask ϱ_{ch}

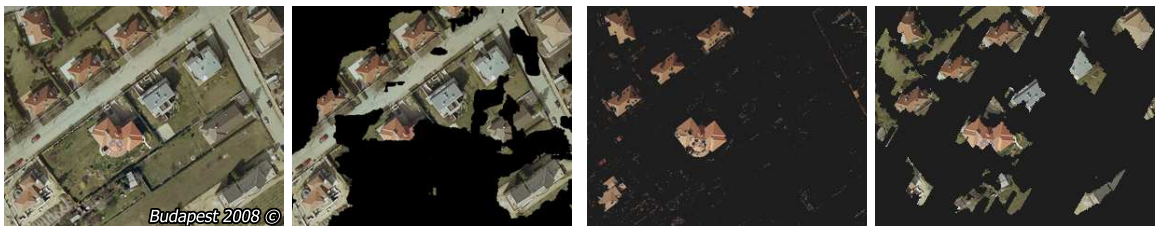
be efficiently used for region based comparison of two remotely sensed images. However, as demonstrated in Fig. 4.1, in cases of high-resolution images with large connected change-regions, a low level change mask cannot efficiently highlight the interesting image content. Stepping up to object level, we develop here a Marked Point Process approach, which models the building population as an optimal configuration of simple geometric objects [42], that is obtained through an iterative process of stochastic birth and death steps (see definitions in Sec. 2.2).

Formally, the input of the proposed method consists of two co-registered aerial or satellite images which were taken from the same area with several months or years of time difference. We consider each building to be constructed from one or many rectangular building segments, and as output we provide the size, position and orientation parameters of the detected building segments, giving information which objects are new, demolished, modified/rebuilt or unchanged [43, 44].

Let us denote by S the common pixel lattice of the input images and by $s \in S$ a single pixel. Let u be a building segment candidate assigned to the input image pair, which is jointly characterized by geometric and temporal attributes. For purposes of dealing with multiple time layers, we assign to each u an index flag, $\xi(u) \in \{1, 2, *\}$, where '$*$' indicates an unchanged object (i.e. present in both images), while '1' and '2' correspond to building segments which appear *only* in the first *or* second image respectively. We will denote the set of all the possible object records u=(c_x, c_y, e_L, e_l, θ, ξ) by $\mathcal{H}$. The output of the proposed model is a configuration of building segments, $\omega \in \mathcal{H}^n$, where n, the number of objects is also unknown.

4.2.2 Feature selection

Since the proposed model obtains the optimal object configuration through stochastic birth-death iterations, two essential questions should be answered based on the image data. First, how can we efficiently generate relevant objects during the *birth* process? Second, how can it be ensured that the adequate objects survive the *death* step? To keep focus on both challenges, we utilize low level and object level features in parallel [52].



(a) Input image (b) Thresh. $P_b^{\mathrm{gr}}(s)$ map (c) Color mask (d) Thresh. $P_b^{\mathrm{sh}}(s)$ map

Figure 4.2: Building candidate regions obtained by the low level (b) gradient (c) color and (d) shadow descriptors

4.2.2.1 Low level features for building detection

We begin the discussion with low level features extracted from individual images. For the purposes of built-in area estimation, at each pixel s we calculate a pair of birth probabilities, $P_b^{(1)}(s)$ and $P_b^{(2)}(s)$, which give the likelihood of s being an object center in image 1, and 2, respectively. The nomination refers to the fact that in the *birth* step the frequency of proposing an object at s will be proportional to the local birth probabilities.

The first feature exploits the fact that regions of buildings should contain edges in perpendicular directions, which can be robustly characterized by *Local Gradient Orientation Density Histograms (GODH)* [139] calculated around each pixel s in a $W_l(s)$ rectangular region. If this region covers a building, the orientation histogram has two peaks, located at 90° degree distance, which can be measured by correlating the histogram with an appropriately matched bi-modal density function. Therafter based on the maximal correlation values we can assign to each pixel s a likelihood $P_b^{\mathrm{gr}}(s)$ that s is center of a building. For the sample image in Fig. 4.2(a), a thresholded P_b^{gr} map is shown in Fig. 4.2(b), marking an estimation for building regions. Furthermore, the location of the peak value of GODH around s estimates the dominant gradient directions in the local neigborhood. Thus, for a building with center s, we expect its θ parameter around a mean orientation value $\mu_\theta(s)$ which is equal to the peak location of the local GODH.

We continue with *roof color filtering*. Several types of roofs can be identified by their typical colors [80]. Let us assume that based on a roof color hypothesis, we extract an indicator mask $\varrho_{\mathrm{co}}(s) \in \{0,1\}$ (e.g. by thresholding a chrominance channel), where $\varrho_{\mathrm{co}}(s) = 1$ marks that s has roof color. Many roof pixels are expected around building centers, thus for each s we calculate the accumulated $\varrho_{\mathrm{co}}-$filling factor in its neighborhood: $\Gamma_s = \sum_{r \in W_l(s)} \varrho_{\mathrm{co}}(r)$. The color birth map value is obtained as $P_b^{\mathrm{co}}(s) = \Gamma_s / \sum_{r \in S} \Gamma_r$. Note that due to color overlapping between the roofs and the background [80], the $\varrho_{\mathrm{co}}(s)$ mask often only contains a part of the building segments, e.g. only *red* roofs are detected in Fig. 4.2(c).

A supplementary evidence for the presence of buildings can be obtained through their *cast shadows* [80, 140]. In several types of remote sensing scenes, a binary shadow mask $\varrho_{\mathrm{sh}}(s)$ can be

derived by filtering pixels from the dark-blue color domain [141]. The relative alignment of shadows to the buildings is determined by the global Sun direction, which can be set with minor user interaction or calculated automatically [80]. Consequently, we can identify the building candidate areas as image regions lying next to the shadow blobs opposing the *Sun direction* (see Fig. 4.2(d) and later Fig. 4.4(b)). As for the shadow based birth map, we use a constant birth rate $P_b^{\mathrm{sh}}(s) = p_0^{\mathrm{sh}}$ within the obtained candidate regions and a smaller constant on the outside.

Up to this point, we have used various descriptors to estimate the location and appearance of the buildings in the individual images. However, a low level change mask ϱ_{ch} demonstrated in Fig. 4.1 can be directly involved in the model, since it separates efficiently the image regions which contain the changed and unchanged buildings, respectively. The probability of change around pixel s is derived as: $P_{\mathrm{ch}}(s) = \sum_{r \in W_l(s)} \varrho_{\mathrm{ch}}(r)/\mathrm{area}\,\{W_l(s)\}$. Considering the change feature, we can exploit an additional information source, which is independent of the object recognizer. During the birth step, we will propose an unchanged object at s with a probability proportional to $(1 - P_{\mathrm{ch}}(s)) \cdot \max_{i \in \{1,2\}} P_b^{(i)}(s)$, while at the same location, the likelihood of generating a changed building segment is $P_{\mathrm{ch}}(s) \cdot P_b^{(i)}(s)$ for image $i \in \{1, 2\}$.

4.2.2.2 Object-Level Features

Besides efficient object generation, the second key point of the applied birth-death dynamics based approach is to validate the proposed building segment candidates. In this section, we construct a $\varphi^{(i)}(u) : \mathcal{H} \to [-1, 1]$ energy function, which calculates a negative building log-likelihood value of object u in the i^{th} image (hereafter we ignore the i superscript). By definition, a rectangle with $\varphi(u) < 0$ is called *attractive* object, and we aim to construct the $\varphi(u)$ function so that attractive objects correspond exclusively to the true buildings.

The process consists of three parts: feature extraction, energy calculation and feature integration. *First*, we define different $f(u) : \mathcal{H} \to \mathbb{R}$ features which evaluate a building hypothesis for u in the image, so that 'high' $f(u)$ values correspond to efficient building candidates. In the *second step*, we construct energy subterms for each feature f, by attempting to satisfy $\varphi_f(u) < 0$ for real objects and $\varphi_f(u) > 0$ for false candidates. For this purpose, we project the feature domain to $[-1, 1]$ with a monotonously decreasing function shown in Fig. 4.3:

$$\varphi_f(u) = \mathcal{M}(f(u), d_0, D) = \begin{cases} \left(1 - \frac{f(u)}{d_0}\right), & \text{if } f(u) < d_0 \\ \exp\left(-\frac{f(u)-d_0}{D}\right) - 1, & \text{if } f(u) \geq d_0 \end{cases} \tag{4.1}$$

Observe that the $\mathcal{M}$ function has two parameters: d_0 and D. While D^f performs data-normalization, d_0^f is the object acceptance threshold concerning feature f: u is attractive according to the $\varphi_f(u)$ term iff $f(u) > d_0^f$.

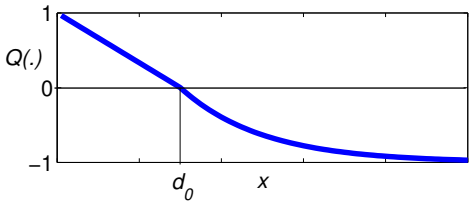
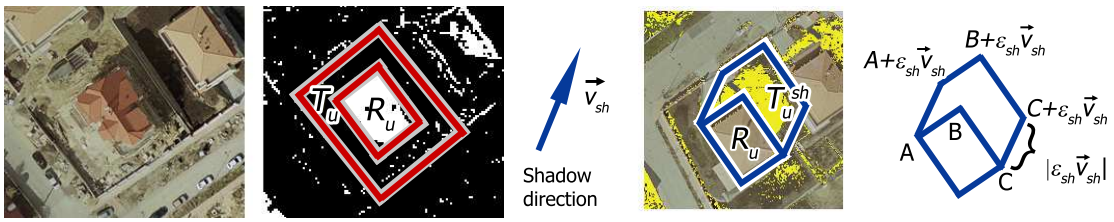


Figure 4.3: Plot of the nonlinear feature domain mapping function $\mathcal{M}(x, d_0, D)$



(a) Color roof feature (b) Shadow feature

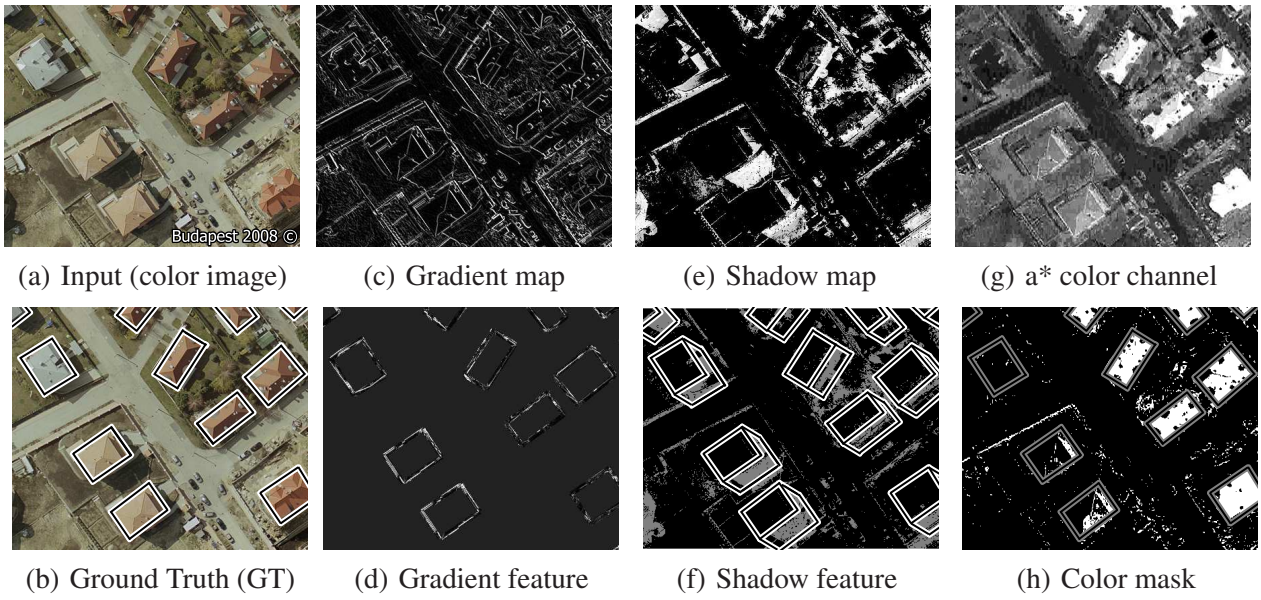
Figure 4.4: Utility of the color roof and shadow features

Finally, we must consider, that the decision based on a single feature f can lead to a *weak classification*, since the buildings and the background may overlap in the f-domain. Therefore, in the *third step*, the joint energy term $\varphi(u)$ must be appropriately constructed from the different $\varphi_f(u)$ feature modules.

We begin the introduction of the different feature models with `gradient` analysis. Below the edges of a relevant rectangle candidate R_u, we expect the magnitudes of the local gradient vectors (∇g_s) to be high and the orientations to be close to the normal vector ($\mathbf{n}_s$) of the closest rectangle side (Fig. 4.5(c)(d)). The $f^{\mathrm{gr}}(u)$ feature is calculated as: $f^{\mathrm{gr}}(u) = \frac{1}{\#\tilde{\partial}R_u} \sum_{s\in\tilde{\partial}R_u} \nabla g_s \cdot \mathbf{n}_s$, where '$\cdot$' denotes scalar product, $\tilde{\partial}R_u$ is the dilated edge mask of rectangle R_u, and $\#\tilde{\partial}R_u$ is the number of pixels in $\tilde{\partial}R_u$. The data-energy term is calculated as: $\varphi_{\mathrm{gr}}(u) = \mathcal{M}(f^{\mathrm{gr}}(u), d^{\mathrm{gr}}, D^{\mathrm{gr}})$.

The calculation of the `roof color` feature is shown in Fig. 4.4(a). We expect the image points to have dominant roof colors inside the building footprint R_u, while the T_u object-neighborhood (see Fig. 4.4(a)) should contain a majority of background pixels. Hence we calculate the $f^{\mathrm{co}}_{\mathrm{in}}(u) = \frac{1}{\#R_u} \sum_{s\in R_u} \varrho_{\mathrm{co}}(s)$ internal and $f^{\mathrm{co}}_{\mathrm{ex}}(u) = f^{\mathrm{co}}_{\mathrm{ex}}(u) = \frac{1}{\#T_u} \sum_{s\in T_u} \left[1 - \varrho_{\mathrm{co}}(s)\right]$ external filling factors, where $\#X$ denotes the area of X in pixels and $\varrho_{\mathrm{co}}(s)$ is the color mask value by s. We prescribe that u should be attractive according to the color term if it is attractive both regarding the internal and external subterms, thus $\varphi_{\mathrm{co}}(u) = \max\left[\mathcal{M}(f^{\mathrm{co}}_{\mathrm{in}}(u), d^{\mathrm{co}}_{\mathrm{in}}, D^{\mathrm{co}}_{\mathrm{in}}), \mathcal{M}(f^{\mathrm{co}}_{\mathrm{ex}}(u), d^{\mathrm{co}}_{\mathrm{ex}}, D^{\mathrm{co}}_{\mathrm{ex}})\right]$.

We continue with the description of the `shadow term`. This step is based on the binary shadow mask $\varrho_{\mathrm{sh}}(s)$, extracted in Sec. 4.2.2.1. Using the *shadow direction* vector $\vec{v}_{\mathrm{sh}}$ (opposite of the Sun direction vector) we identify a shadow candidate area T^{sh}_u next to the R_u object region, as demonstrated in Fig. 4.4(b). Thereafter, similarly to the color feature, we expect low shadow pres-



(a) Input (color image) (c) Gradient map (e) Shadow map (g) a* color channel

(b) Ground Truth (GT) (d) Gradient feature (f) Shadow feature (h) Color mask

Figure 4.5: Illustration of the feature maps in the BUDAPEST 2008 image. Gradient and shadow features are relevant in the left-bottom regions, while the color descriptor is efficient in the top-right image parts. In image (d), the gradient feature is shown under the GT object borders.

ence in the R_u internal and a high one in the T_u^{sh} external region, which constraints are represented by $f_{\mathrm{in}}^{\mathrm{sh}}(u) = \frac{1}{\#R_u} \sum_{s \in R_u} \left[1 - \varrho_{\mathrm{sh}}(s)\right]$ and $f_{\mathrm{ex}}^{\mathrm{sh}}(u) = \frac{1}{\#T_u^{\mathrm{sh}}} \sum_{s \in T_u^{\mathrm{sh}}} \varrho_{\mathrm{sh}}(s)$ features. The energy term is obtained as: $\varphi_{\mathrm{sh}}(u) = \max\left[\mathcal{M}(f_{\mathrm{in}}^{\mathrm{sh}}(u), d_{\mathrm{in}}^{\mathrm{sh}}, D_{\mathrm{in}}^{\mathrm{sh}}), \mathcal{M}(f_{\mathrm{ex}}^{\mathrm{sh}}(u), d_{\mathrm{ex}}^{\mathrm{sh}}, D_{\mathrm{ex}}^{\mathrm{sh}})\right]$. Note that this approach does not require accurate building height information, since we do not penalize it, if shadow blobs of long buildings exceed the T_u^{sh} regions.

The next step is *feature integration*. Since as shown in Fig. 4.5, individual features may be in themselves inappropriate for modeling complex scenes, the proposed framework enables flexible *feature integration*. From the feature primitive terms introduced in Sec. 4.2.2.2, first we construct building prototypes. For each prototype we can prescribe the fulfillment of one or many feature constraints whose φ_f-subterms are connected with the $\max$ operator in the joint energy term of the prototype (logical AND in the negative log-likehood domain).

Additionally, several building prototypes can be detected simultaneously in a given image pair, if the prototype-energies are joined with the $\min$ (logical OR) operator. Thus the final object energy term is derived by a logical function, which expresses some prior knowledge about the image and the scene, and it is chosen on a case-by-case basis. For example, in the BUDAPEST pair we use two prototypes: the first one prescribes the edge and shadow constraints, the second one the roof color, thus the joint energy is calculated as: $\varphi(u) = \min\left\{\max\left\{\varphi_{\mathrm{gr}}(u), \varphi_{\mathrm{sh}}(u)\right\}, \varphi_{\mathrm{co}}(u)\right\}$. Similarly, for the BEIJING images (see Fig. 4.6, bottom) we use gradient (φ_{gr}) & shadow (φ_{sh}) and homogeneity based (detailed in [10]) prototypes.

4.2.3 Multitemporal MPP configuration model and optimization

In this section we represent the building change detection task as an energy minimization problem. Following the definitions from Sec. 2.2.1, u denotes a given building segment, and the goal of the proposed approach is to extract an $\omega = \{u_1, \ldots, u_n\} \in \Omega$ object configuration, where n, the number of building segments in initially unknown. The object neighborhood is defined here in a straightforward way: we say that $u \sim v$ if their rectangles R_u and R_v intersect. By denoting with $\mathcal{F}$ the union of all image features derived from the input data, the goal is to minimize a classical MPP energy function of (2.21):

$$\Phi_{\mathcal{F}}(\omega) = \sum_{u \in \omega} A_{\mathcal{F}}(u) + \gamma \cdot \sum_{\substack{u,v \in \omega \\ u \sim v}} I(u,v) \tag{4.2}$$

Here $A_{\mathcal{F}}(u) \in [-1,1]$ and $I\,(u,v) \in [0,1]$ are the data dependent unary and the prior interaction potentials, respectively, and $\gamma > 0$ is a weighting factor between the two energy terms. Thus the Maximum Likelihood (ML) configuration estimate can be calculated as $\omega_{\mathrm{ML}} = \mathrm{argmin}_{\omega \in \Omega} \left[\Phi_{\mathcal{F}}(\omega)\right]$.

Unary potentials characterize a given building segment candidate $u = \{c_x, c_y, e_L, e_l, \theta, \xi\}$ as a function of the local image data in both images, but independently of other objects of the population. This term encapsulates the building energies $\varphi^{(1)}(u)$ and $\varphi^{(2)}(u)$ extracted from the 1^{st}, resp. 2^{nd}, image (Sec. 4.2.2.2) and the low level similarity information between the two time layers which is described by the $\varrho_{\mathrm{ch}}(.)$ change mask (Sec. 4.2.2.1).

We remind the Reader that our approach marks each building segment u with an image index flag from the set $\{1, 2, *\}$, depending on that u appears in one [$\xi(u) \in \{1,2\}$] or both [$\xi(u) = *$] of the input images. In this way, the classification of the building segment u is straightforward: u is *unchanged* iff $\xi(u) = *$; *new* iff $\xi(u) = 2$ and $\nexists v \in \omega : \{\xi(v) = 1$, u and v overlap$\}$; and *demolished* iff $\xi(u) = 1$ and $\nexists v \in \omega : \{\xi(v) = 2$, u and v overlap$\}$. Modified buildings are considered as two objects u_1 and u_2, so that $\xi(u_1) = 1$, $\xi(u_2) = 2$.

Three *soft constraints* are considered by the potential terms in the various cases. *First*, for an unchanged building u, we expect low object energies in both images, and penalize textural differences (pixels with $\varrho_{\mathrm{ch}}(s) = 1$) under its footprint R_u. *Second*, for a demolished or modified building in the first image, we expect low $\varphi^{(1)}(u)$, $\varphi^{(2)}(u)$ is indifferent, but we penalize high similarity under the footprint. *Third*, for a new or modified building in the second image, we expect low $\varphi^{(2)}(u)$, $\varphi^{(1)}(u)$ is indifferent, while high local similarity is penalized again.

Consequently, using the $\mathbf{1}\{E\} \in \{0,1\}$ indicator function for an event E, the $A_{\mathcal{F}}(u)$ potential is calculated as:

$$\begin{aligned} A_{\mathcal{F}}(u) = {} & \mathbf{1}\left\{\xi(u) \in \{1,*\}\right\} \cdot \varphi^{(1)}(u) + \mathbf{1}\left\{\xi(u) \in \{2,*\}\right\} \cdot \varphi^{(2)}(u) + \\ & + \mathbf{1}\left\{\xi(u) = *\right\} \cdot \frac{1}{\#R_u} \sum_{s \in R_u} \varrho_{\mathrm{ch}}(s) + \mathbf{1}\left\{\xi(u) \in \{1,2\}\right\} \cdot \frac{1}{\#R_u} \sum_{s \in R_u} \big(1 - \varrho_{\mathrm{ch}}(s)\big) \end{aligned} \tag{4.3}$$

On the other hand, *interaction* potentials realize prior geometrical constraints: they penalize intersection between different object rectangles sharing the time layer (see earlier Fig. 2.6):

$$I(u,v) = \mathbf{1}\left\{\xi(u) \simeq \xi(v)\right\} \cdot \frac{\#(R_u \cap R_v)}{\#(R_u \cup R_v)} \tag{4.4}$$

where $\xi(u) \simeq \xi(v)$ relation holds iff $\xi(u) = \xi(v)$, or $\xi(u) = *$, or $\xi(v) = *$. Since $\forall u, v : I(u,v) \geq 0$, the optimal population should exclusively consist of objects with negative data terms (i.e. attractive objects): if $A_{\mathcal{F}}(u) > 0$, removing u from the configuration results in a lower $\Phi_{\mathcal{F}}(\omega)$ global energy (4.2). Note also that according to eq. (4.2), the interaction term plays a crucial role by penalizing multiple attractive objects in the same or strongly overlapping positions.

By fixing the $A_{\mathcal{F}}(u)$ and $I(u,v)$ potential terms, the $\Phi_{\mathcal{F}}(\omega)$ configuration energy is completely defined, and the optimal ω_{ML} building population can be obtained by minimizing eq. (4.2). For this purpose, we have developed the *bi-layer Multiple Birth and Death* (bMBD) algorithm, which is presented in Appendix C. The bMBD method extends the conventional MBD technique by handling two time layers, thus it encapsulates change and object information simultaneously. Pairs of consecutive birth and death processes are iterated until convergence is obtained in the global configuration. In the *birth* step, multiple object candidates are generated randomly according to the birth maps $P_b^{(i)}(s)$ in time layers $i \in \{1,2\}$, and as a further novelty, also considering the change probabilities $P_{\mathrm{ch}}(s)$ and the expected parameter maps such as $\mu_\theta^{(i)}(s)$ $i \in \{1,2\}$. The *death* process attempts to eliminate weak objects based on the global configuration energy.

4.2.4 Experimental study of the mMPP model

During the evaluation, we validated the three key developments of the proposed model, with a comparison to the state-of-the-art: (i) the proposed multiple feature based building appearance model, (ii) the joint object-change modeling framework and (iii) the non-homogeneous object birth process based on low level features. We have evaluated our method using eight significantly different sets of aerial and satellite image pairs, published as the *SZTAKI-INRIA Building Detection Benchmark Set*[1]. For parameter settings, we have chosen in each data set 2-8 buildings ($\approx 5\%$) as training data, while the remaining Ground Truth labels have only been used to validate the detection results. Qualitative results are shown in Fig. 4.6, and in Appendix C (Fig. C.1).

We perform quantitative evaluation both at object and pixel levels. At the object level, we first need to establish a non-ambiguous assignment between the detected objects and the GT object samples. As a similarity feature, we use the normalized intersection area between the object figures, and we find the optimal match between the configuration elements with the Hungarian Algorithm (HA) [9, 142]. A detected object is labeled as True Positive (TP), if the HA matches

[1]Url: `http://mplab.sztaki.hu/remotesensing/building_benchmark.html`

Table 4.1: Numerical object level and pixel level comparison of the SIFT, Gabor, EV, SM and the proposed methods (MPP) on each test data set (best results in each row are typeset by bold.)

Dataset		Object level performance										Pixel level performance					
		SIFT		Gabor		EV		SM		**MPP**		EV		SM		**MPP**	
	#o.	FN	FP	FN	FP	FN	FP	PN	FP	FN	FP	Pr	Rc	Pr	Rc	Pr	Rc
Bp	41	20	10	8	17	11	5	9	**1**	**2**	4	73	46	**84**	61	82	**71**
An	21	8	5	**0**	1	2	**0**	2	1	1	**0**	**91**	73	84	**79**	83	74
Bg	17	7	2	9	8	2	3	4	2	**1**	**0**	59	26	71	**72**	**93**	71
Sz	57	17	26	17	23	10	18	11	5	**4**	**1**	61	62	79	71	**93**	**75**
Cd	123	55	9	12	24	14	20	20	25	**5**	**4**	73	51	75	61	**83**	**69**
Bs	80	34	9	32	8	11	13	18	15	**7**	**6**	56	30	59	41	**73**	**51**
Nm	152	69	14	24	14	**18**	32	30	58	**18**	**1**	60	32	62	55	**78**	**60**
Mc	171	NA	NA	53	85	46	17	53	42	**19**	**6**	64	38	60	56	**86**	**63**
F-s%		66.3		79.9		84.2		79.8		**94.4**		53.7		66.8		**74.3**	

Datasets: Budapest (Bp), Abidjan (An), Beijing (Bg), Szada (Sz), Cot d'Azur (Cd), Bodensee (Bs), Normandy (Nm) and Manchester (Mc), #o is the number of objects

it to a GT object with an overlapping rate of more than r_h (used $r_h = 10\%$). Unpaired detection samples are marked as False Positive (FP), unpaired GT objects as False Negative (FN) hits. For change detection evaluation, we also count missing and false change alarms (MC, FC).

At pixel level we investigate how accurate the extracted object outlines are: we compare the resulting building footprint masks to the Ground Truth mask, and calculate the Precision (Pr) and Recall (Rc) values of the pixel level detection. Finally, the F-score (F-s), taken as the harmonic mean of Pr and Rc, can be calculated both at object and at pixel levels.

By evaluation of the building detection component, we presented numerical and qualitative comparison results versus four single-view building detection techniques, called SIFT [143], Gabor [144], Edge Verification (EV) [80], and the Segment-Merge (SM) model [78]. Quantitative evaluation results are shown in Table 4.1. Since SIFT and Gabor extract the building centers instead of estimating the outline, they are only involved in the object level comparison. Numerical results confirm that the proposed model surpasses all references with 10-26% at object level and with 5-18% at pixel level. According to our analysis, the improvements are particularly related to two key properties: the stochastic object generation process and the parallel utilization of multiple features in the building description module. In terms of computational complexity, processing 1MPixel images with the MPP model takes in average less than 1 minute. The proposed approach is competitive with most reference techniques regarding the running time parameter, as detailed experiments in [10] confirm.

After testing the introduced building detector module in single images, we continue with the

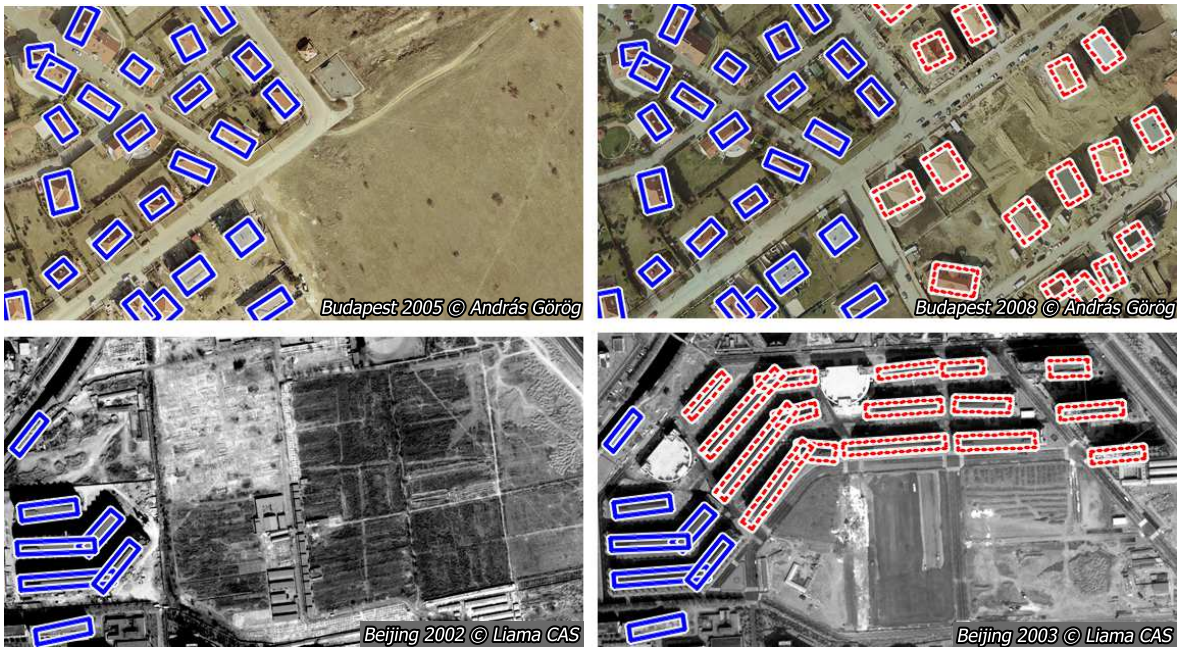


Figure 4.6: Results on BUDAPEST (top, image part - provider: András Görög) and BEIJING (bottom, provider: Liama Laboratory CAS, China) image pairs, marking the unchanged (solid rectangles) and changed (dashed) objects

validation of the proposed joint object-change classification framework. We compared the mMPP method to a conventional Post Detection Comparison (PDC) [145] technique, where the buildings are separately extracted from the two image layers, and the change information is a posteriori estimated through comparing the location, geometry and spectral characteristics of the detected objects. Experiments on the Budapest, Abidjan, Beijing, Szada datasets have shown that 90% of false change alarms of PDC were eliminated by mMPP (see details in Table C.1 of Appendix C).

To demonstrate the advantages of the *Feature Based Birth Process* (FBB), we compared the convergence speed of the bMBD optimization using the proposed FBB and the conventional Uniform Birth (UB) processes. In the UB case, the $P_b^{(i)}(s)$ and $P_{\mathrm{ch}}(s)$ maps follow uniform distributions and the orientation parameters are also set as uniform random values. We experienced that the FBB approach reaches the final error rate with three times less birth calls than the UB. Moreover, using the UB process the pixel level accuracy rates converge much slower than the object errors; to reach the 75% pixel level F-score, we need to generate $400,000$ objects with the UB map, and only $24,000$ building candidates with the proposed FBB map.

4.3 A point process model for target sequence analysis

While the mMPP model introduced in the previous section provides a solution for object level change detection between two remotely sensed images, a significantly different problem family corresponds to scenarios, where a moving target must be followed across several frames of a measurement sequence. In this section, we propose a novel framework for object level time sequence analyis, which we call henceforward Multiframe MPP (F^{m}MPP).

The F^{m}MPP framework simultaneously considers the consistency of the observed data and the fitted objects in the individual time instances of the measurement, and also exploits interaction constraints between the object parameters in the consecutive frames of the sequence. We should point out, that the optimization step is a particularly critical issue in the multiframe scenario: the dimension of the target sequence's parameter space may be very large, as it is proportional to the number of frames. For this reason, in the proposed model we merge the advantages of both the bottom-up and inverse approaches (see the definitions from Chapter 1). First, we apply a *bottom-up* detector for initial target extraction, which processes the sequence frame-by-frame. This step is quick, but we must expect that the results are notably poor in low quality frames. The output of the bottom-up detector provides the initial state of the F^{m}MPP optimization process, which yields the final output ensuring permanent target structure and smooth motion over the sequence via inter-frame constraints.

4.3.1 Application on moving target analysis in ISAR image sequences

We introduce the F^{m}MPP approach in the application context of moving target analysis in airborne Inverse Synthetic Aperture Radar (ISAR) image sequences. Remotely sensed ISAR images can provide valuable information for target classification and recognition in several difficult situations, where optical [13] or SAR imaging techniques fail [146, 147]. However, robust feature extraction and feature tracking in the ISAR images are usually difficult tasks due to strong image noise and lack of available details about the structure of the imaged targets, which artifacts can lead to significant detection errors in several low quality frames [38]. Some previous studies have proposed frame selection strategies to exclude low quality frames from the analysis. However, as pointed out in [148] extracting reliable features for frame selection may often fail. On the other hand, assuming that the target has a fixed size and structure; and small displacement is expected between consecutive time appearances, inter-frame information can be exploited to refine the detection procedure. For this reason, our proposed system does not drop any frames of the input sequence, but it implements an approach where the detection result on the actual frame jointly depends on the current image data and the neighboring frame's target parameters.

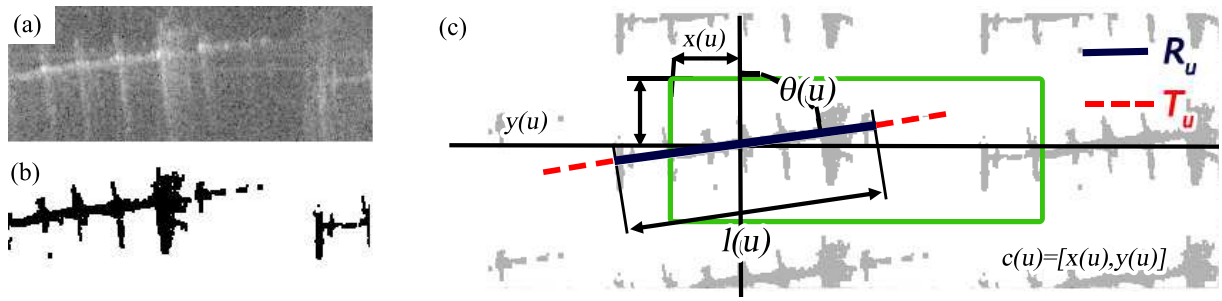


Figure 4.7: Target representation in an ISAR image: (a) input image with a single *ship* object (b) binarized image (c) duplicated image and target fitting parameters. Original image border is shown by the green rectangle

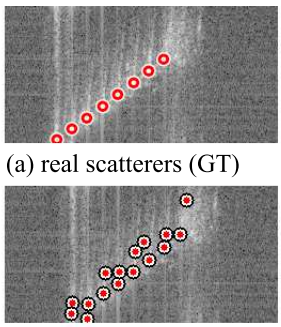
Besides the length and axis line extraction of the target scatterer, another issue is to detect characteristic features of the objects which provide relevant information for the identification process. For this purpose, we identify permanent bright points in the imaged targets, which are produced by stronger scatterer responses from the illuminated objects (see Fig. 4.8(a)). However, due to the presence of speckle, image defocus and scatterer scintillation, a significant number of missing and false scatterer-like artifacts appear in the individual frames, thus we focus on their elimination with spatio-temporal filtering constraints.

The contributions of the section are twofold. On one hand, we introduce the general Multiframe (F^{m}MPP) MPP framework, which provides a novel Bayesian tool for time sequence analysis in remotely sensed scenarios. On the other hand, we propose a concrete implementation of the F^{m}MPP method on the analysis of large carrier ships and airplanes from ISAR data, and perform a detailed quantitative validation on a real data set, which contains eight ISAR image sequences with 545 manually evaluated frames.

4.3.2 Problem definition and notations

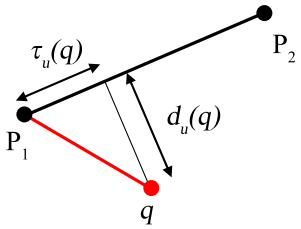
The input of the proposed algorithm is an n-frame long sequence of 2D ISAR data, imaged in the *Range-Doppler* domain, which contains a single *ship* (or *airplane*) target. Let us denote by S the joint pixel lattice of the images, and by $s \in S$ a single pixel. The normalized log-amplitude of pixel s in frame $t \in \{1, 2, \ldots, n\}$ is marked with $g_t(s)$. The logarithmic image representation suits well the widely adopted log-normal statistical models of ISAR target segmentation [149].

In the following, we denote by u_t a target candidate in frame t. Each target's axis line segment is described by the $c(u) = [x(u), y(u)]$ center pixel $l(u)$ length and $\theta(u)$ orientation parameters (see Fig. 4.7(c)). In addition, an initially unknown $K(u)(\leq K_{\max})$ number of scatterers can be assigned to the targets, where each scatterer q_i is described in the target line segment's coordinate system by the relative line directional position, $\tau_u(q_i)$, and the signed distance, $d_u(q_i)$ from the



(a) real scatterers (GT)

(b) LocMax filter results



P_1, P_2: end points of the line segment
q: scatter candidate
$\tau_u(q)$: normalized line directional position (fit into the [0,1] interval)
$d_u(q)$: signed distance from the P_1P_2 line segm. (measured in pixels)

(c) Parameterization of the scatterer position

Figure 4.8: Dominant scatterer detection problem: (a) highlighted true scatterers, i.e. Ground Truth (GT), (b) LocMax filter result, (c) parameterization

center line of the parent object u (see Fig. 4.8(c)). The goal is to obtain a $\omega = \{u_1, u_2, \ldots, u_n\}$ target sequence, which we call *configuration* in the following.

4.3.3 Data preprocessing in a bottom-up approach

Data preprocessing consist of four consecutive steps: foreground-background segmentation, initial center alignment and line segment estimation, scatterer candidate set extraction, and scatterer filtering.

In the first step, we segment the ISAR images into *foreground* and *background* classes by a binary Markov Random Field (MRF) model [7] to decrease the spurious effects of speckle noise. (See Appendix C, Sec. C.2.1 for details.)

Thereafter, to get an *initial estimation of the target axis segment*, we detect first the axis line using the Hough transform of the foreground mask. At this point, we also have to deal with a problem which originates from the ISAR image synthesis module. The image formation process considers the images to be spatially periodic both in the horizontal and vertical directions, then, the imaging step estimates the target center, and attempts to crop the appropriate Rectangle of Interest (ROI) from this periodic image (a correctly cropped frame is in Fig. 4.8(a)). However, if the center of the ROI is erroneously identified, the target line segment may 'break' into two (or four) pieces, which case appears in Fig. 4.7(a). Therefore, in the proposed image processing approach, we search for the longest foreground segment of the axis line in a duplicated mosaic image, which step also re-estimates the center of the input frame (see Fig. 4.7(c)).

Scatterer candidate extraction exploits the fact that permanent scatterers cause dominantly high amplitudes in the ISAR images; however the amplitudes may significantly vary over the consecutive frames, and we must expect notable differences between different scatterers of the same frame.

In our implementation, the Local Maxima (LocMax) filter is applied to extract the Preliminary Scatterer Candidates (output results are shown in Fig. 4.9(a)). Thereafter, we propose an iterative Random sample consensus (RANSAC) based solution to discriminate the real scatterers from the false candidates, with utilizing the temporal persistence of the scatterer positions and the line-structure of the imaged targets [33] (see outputs in Fig. 4.9(b)).

4.3.4 Multiframe Marked Point Process Model

In this section, we introduce the Multiframe Marked Point Process model, which enables to characterize whole target sequences instead of individual objects, through exploiting information from entity interactions. Following the classical Markovian approach, each target sample may only affect objects in its *neighboring frames* directly, using a ζ-radius frame neighborhood. Using again the $\mathcal{F}$ notation for the union of all image features derived from the input data, we characterize a given ω target sequence with a $\Phi_{\mathcal{F}}(\omega)$ data-driven Gibbs energy function:

$$\Phi_{\mathcal{F}}(\omega) = \sum_{t=1}^{n} A_{\mathcal{F}}(u_t) + \gamma \cdot \sum_{t=1}^{n} I(u_t, \omega_t) \tag{4.5}$$

As it appears in the above formula, $\Phi_{\mathcal{F}}(\omega)$ consists of a data dependent term, $A_{\mathcal{F}}(u_t) \in [-1, 1]$ called the unary potential, and a prior term $I(u_t, \omega_t) \in [0, 1]$, called the interaction potential, where $\omega_t = \{u_{t-\zeta}, \ldots, u_t, \ldots, u_{t+\zeta}\}$ is a sub-sequence of u_t's 2ζ-nearest neighbors. Parameter γ is a positive weighting factor between the two potential terms.

The $A_{\mathcal{F}}(u_t)$ unary potential is composed of two parts:

$$A_{\mathcal{F}}(u_t) = \frac{1}{2}\left(A^{\mathrm{B}}_{\mathcal{F}}(u_t) + A^{\mathrm{Sc}}_{\mathcal{F}}(u_t)\right)$$

where $A^{\mathrm{B}}_{\mathcal{F}}(u_t)$ is the *body-term* and is the $A^{\mathrm{Sc}}_{\mathcal{F}}(u_t)$ *scatterer-term*.

The *body-term*, is based on a *body fitting feature* $f_b(u)$, which favors object candidates, where under the line segments R_u we find in majority foreground classified pixels in the actual frame, while the neighboring outside area T_u covers background regions (see Fig. 4.7). Thereafter, $A^{B}_{\mathcal{F}}(u) = Q(f_b(u), d_0, 10)$, where the monotonously decreasing $\mathcal{M}$ function defined by eq. (4.1) is utilized again. Parameter d_0 is used as acceptance threshold for valid objects.

On the other hand, the *scatterer-term* penalizes scatterers that are not located at local Maxima of the ISAR image: $A^{\mathrm{Sc}}_{\mathcal{F}}(u) = \mathcal{M}\left(\frac{1}{K(u)} \cdot \sum_{i=1}^{K(u)} \Psi(i, u),\ d_\Psi\right)$, where $\Psi(i, u) = 0$ if q_i is in a local maximum in the input ISAR frame, $\Psi(i, u) = 1$ otherwise. Parameters d_0 and d_Ψ are set by training samples.

Interaction potentials are responsible for involving temporal information and prior geometric knowledge in the model. Since the observed object's structure can be considered rigid, we usually

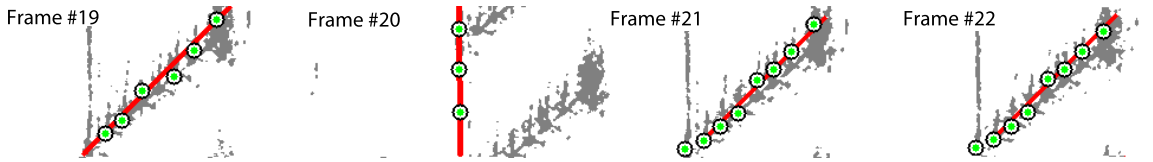
experience strong correlation between the target parameters in the consecutive frames. Since due to the imaging technique, the $c(u)$ center is not relevant regarding the real target position, we only penalize high differences between the $\theta(u)$ angle and $l(u)$ length parameters, and significant differences in the normalized scatterer positions and scatterer numbers between close-in-time images of the sequence. The prior interaction term $I(u_t, \omega_t)$ is constructed as the weighted sum of four subterms: the median length difference $I_l(u_t, \omega_t)$, the median angle difference $I_\theta(u_t, \omega_t)$, the median scatterer number difference $I_{\#\mathrm{s}}(u_t, \omega_t)$ and the median scatterer alignment difference $I_{\mathrm{sd}}(u_t, \omega_t)$. Here the first three sub-terms are calculated as the median values of the parameter differences between the actual and the nearby frames, while the scatterer alignment difference feature evaluates the similarity of the relative scatterer positions on the objects of close frames [7].

4.3.5 Multiframe MPP optimization

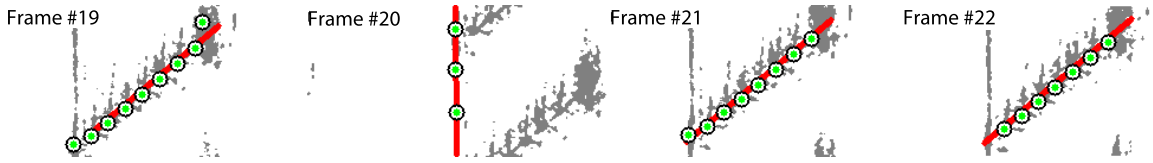
As mentioned in the beginning of this section, our proposed optimizer solution is initialized with the output of the preliminary detector of Sec. 4.3.3, which provides an initial configuration which is in most of the frames consistent with the input data. Thereafter, we proceed to an iterative refinement algorithm, which scans in each step the whole sequence, and attempts to replace the actual objects with more efficient ones considering the data and prior constraints in parallel. The two key points of this procedure are (i) the generation of new object candidates and (ii) the evaluation of the proposed objects w.r.t. the current configuration and the input data.

For object generation, we use two types of moves: a Perturbation Kernel and a RANSAC based birth kernel, which are chosen randomly at each step of each iteration. The Perturbation kernel clones the actual object either from the current, or the previous or the next frame; and it adds zero mean Gaussian random values to the center position, length and orientation parameters. Finally, the scatterer positions are cloned from the object of the current frame and optionally new scatterers are added or some scatterers are removed. The RANSAC based birth kernel re-estimates the optimal line according to the preliminary scatterer candidates with the RANSAC algorithm. The pseudo codes of these functions are provided in Appendix C (Algorithm C.2).

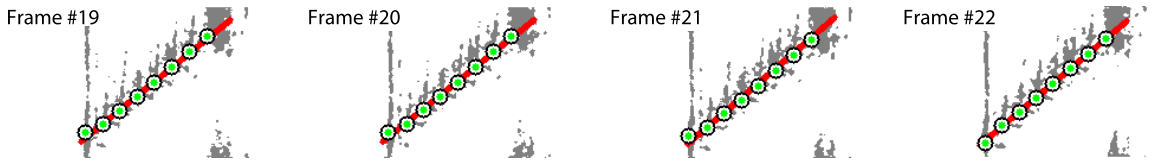
The proposed optimization algorithm iterates object proposal and evaluation steps, which are followed by the possible replacements of the original objects versus newly generated ones. Let us assume that we are currently in the kth iteration of the process. To decide if we accept or decline the replacement of the object on the tth frame for the newly proposed object, u, we calculate first the energy difference $\Delta\Phi_\omega(u, t)$ between $\omega^{[k]}$, the original configuration before the kth iteration, and the configuration ω^* we would get from $\omega^{[k]}$ by replacing $u_t^{[k]}$ by u. It is important to note that to derive the energy difference we should only examine the objects in the ζ-neighborhood of frame t and calculate the concerning unary and interaction potential terms. $\Delta\Phi_\omega(u, t) < 0$ means that the move results in decreasing global energy level. However, to prevent us from finishing the



(a) Initial detection results (Preprocessing, first step)



(b) RANSAC-based refinement (Preproc., second step)



(c) Final F^{m}MPP output after the iterative optimization

Figure 4.9: Center alignment and target line extraction results on Frames #19-22 of the `SHIP1` ISAR image sequence. Top: initial detection Middle: RANSAC re-estimation Bottom: proposed F^{m}MPP model.

algorithm too early in a low quality local energy minimum, we embed the iterative process into a simulated annealing framework. In this way, as a function of the $\Delta\Phi_\omega(u,t)$ energy difference, we calculate a probability value of accepting the replacement move, and the decision is done by a random choice based on this probability. Regarding the cooling scheme, we have followed the implementation of [87]. Details of optimization are provided in Algorithm C.3 of Appendix C. Final detection results on the previously discussed sample frames are shown in Fig. 4.9(c)).

4.3.6 Experimental results on target sequence analysis

We have tested our method on seven airborne ISAR image sequences about different ship targets (Fig. 4.10). The ship data set contains 520 evaluated ISAR frames (40 to 90 frames have been evaluated in each sequence) and 4250 true scatterer appearances (8 or 9 scatterers in each frame). For quantitative validation, we have manually created Ground Truth (GT) data for both the axis segments and the scatterer positions (for longer sequences we have evaluated the first 90 frames).

To consider different evaluation aspects, we have defined three error measures. The *Normalized Axis Parameter Error* (E_{AX}) is calculated as the sum of the x-y center position and axis length errors normalized with the length of the GT target, and the angle error normalized by 90°. The *Scatterer Detection Rates* characterize the correctness of permanent scatterer identification. First, the corresponding detected and GT scatterers are automatically matched to each other by the Hungarian algorithm [142], based on the $\tau(q)$ parameter. A match is only considered valid if the

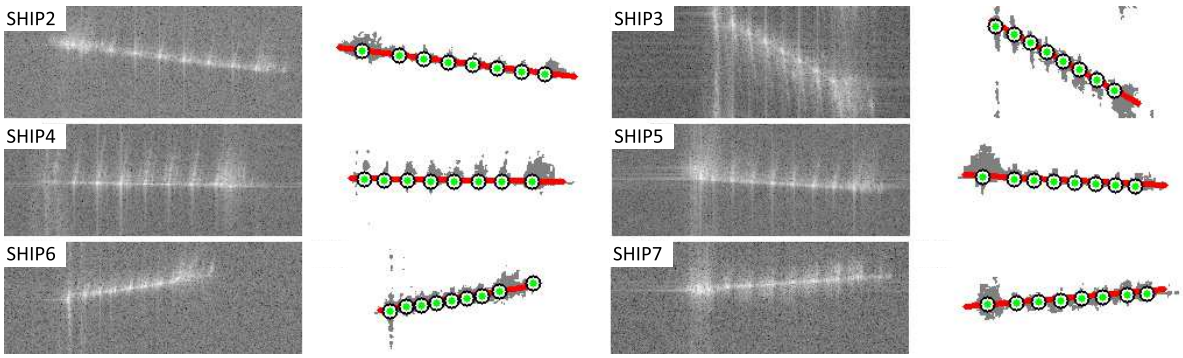


Figure 4.10: Sample frames from the SHIP2-SHIP7 data sets, and the corresponding detection results of the F^{m}MPP approach obtained by the optimization of the proposed ISAR sequence based model.

distance of the assigned feature points is lower than a threshold. Thereafter, we count the number of true positive (TP), false negative (FN) and false positive (FP) scatterers. The third feature is the *Average Scatterer Position Error* (E_{SP}), which is measured in pixels. Note that TP should be large for an efficient solution, while all other parameters should be low.

Similarly to our qualitative experiences (Fig. 4.9), we have observed during the quantitative evaluation that the detection qualities have significantly improved over the three consecutive steps of the proposed algorithm. While after the *Initial* detection step, we measured in average over the seven test sequences $E_{\mathrm{AX}} = 10.8$, TP$=$ 563, FP$=$ 63, FN$=$ 44, and $E_{\mathrm{SP}} = 2.9$ values, following *RANSAC*-based refinement we obtained $E_{\mathrm{AX}} = 7.6$, TP$=$ 581, FP$=$ 49, FN$=$ 25, and $E_{\mathrm{SP}} = 1.5$, while the final F^{m}MPP energy optimization yielded $E_{\mathrm{AX}} = 3.8$, TP$=$ 587, FP$=$ 20, FN$=$ 19, and $E_{\mathrm{SP}} = 0.87$. Detailed results with respect to all error parameters for each test set are shown in Appendix C.

The proposed model can be generalized to analyze various targets in ISAR image sequences. For example, airplanes appear as cross-like structures in the ISAR image sequences, where at least one of the wings can be clearly observed. Apart from the length and orientation of the body axis segment, the length of the wings and their connecting positions to the airplane body are also relevant shape parameters. For this reason, we use a cross shaped airplane model, as shown in Fig. 4.11. Similarly to the ship detection procedure, the airplane extraction process consists of a coarse preliminary detection step, and the F^{m}MPP based iterative refinement step. The preliminary detection starts with the extraction of the body line, using the same Hough transform based technique as introduced for the ship detection process. Secondly, the initial wing root position is obtained with exhaustive search by histograming the silhouette pixels, which can be perpendicularly projected to the same points of the body line. In the F^{m}MPP based refinement stage the $A_{\mathcal{F}}(u_t)$ data term is calculated in an analogous manner to the ship model, the difference is that the filling factors for the left and right wings are separately calculated, and their minimum (i.e. the better one) counts into

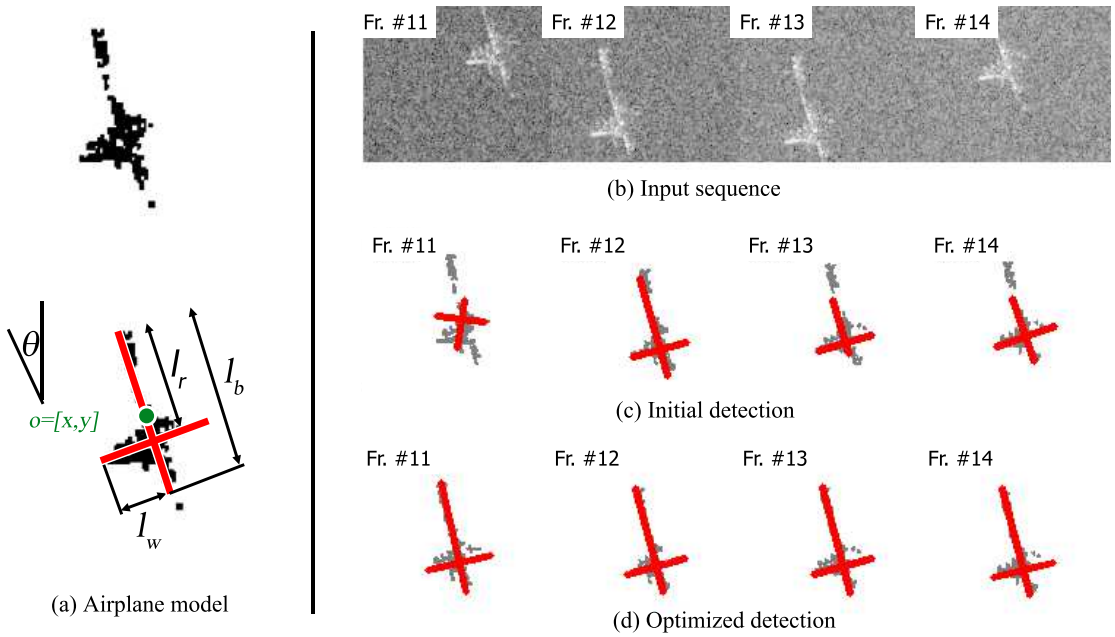


Figure 4.11: Airplane detection example: (a) Airplane silhouette and the cross shaped fitted model (b)-(d) Comparing the results of the *initial* and the optimized F^{m}MPP detection in four sample frames

the data term of the model. This later feature is necessary, since usually only one of the wings is fully visible in the ISAR data. Results of the airplane detection for four sample frames are demonstrated in Fig. 4.11 showing the output at the preliminary stage and after the F^{m}MPP optimization. Similar improvement can be observed to the ship detection scenarios. Note that it is also often possible to observe permanent scatterers in images of airplane targets. However, since airplane scatterers can appear both in the wings and in the body, their geometric alignment patterns may be more complex than in cases of the linear vessels.

The processing speed of the proposed algorithm varies over the different test sets between 2 frames per second (fps) and 5fps, since the computational complexity depends on various factors, such as length of the sequence, image size, target length, quality of the initial detection step and number of scatterers. In most cases, the computational overload of the iterative optimization is not significantly higher than the cost of the initialization and the RANSAC steps.

4.4 Parameter settings in dynamic MPP models

We can divide the parameters of the proposed dynamic MPP methods into three groups corresponding to the *prior model*, *data model* and the *optimization*.

The *prior model* parameters of the mMPP model, such as the l window size for GODF calculation by (see Sec. 4.2.2.1) and maximal/minimal rectangle side lengths at the difference scales,

depend on image resolution and expected object dimensions, thus they are set based on sample objects. As for the F^{m}MPP approach the r_{fg}, T_τ, $d^l_{\max}$, $d^\theta_{\max}$ $d^K_{\max}$ and $d^{\mathrm{sd}}_{\max}$ factors are also calibrated in a supervised way. Here further relevant prior parameters are the weighting factors within the $I(u_t, \omega_t)$ interaction term, we used here uniform weights $\delta_l = \delta_\theta = \delta_{\#s} = \delta_{\mathrm{sd}} = 0.25$. We used a constant $\gamma = 2$ weight between the data term and the overlapping coefficient in eq. (4.2) and (4.5).

The parameters of the *data model* in mMPP are estimated based on training image regions containing *Ground Truth* objects $\{u^{\mathrm{gt}}_1, u^{\mathrm{gt}}_2, \ldots, u^{\mathrm{gt}}_n\}$. Consider an arbitrary $f(u)$ feature from the feature library (e.g. $f^{\mathrm{gr}}(u)$ gradient descriptor for building detection). We remind the Reader that each $f(u)$ of our model is a noisy quality measure and the corresponding energy term is obtained as $\varphi_f(u) = \mathcal{M}(f(u), d^f_0, D^f)$ (see Sec. 4.2.2.2). Here we set the normalizing constant as $D^f = \max_j f(u^{\mathrm{gt}}_j) - \min_j f(u^{\mathrm{gt}}_j)$. Exploiting that the $\mathcal{M}$ transfer function is monotonously decreasing with a sole root $f(u) = d^f_0$, object u is attractive in image i (i.e. $\varphi^{(i)}_f(u) < 0$) iff $f(u) > d^f_0$. Consequently, increasing d^f_0 may decrease the false alarm rate and increase the missing alarms corresponding to the selected feature. Since in the proposed model we can simultaneously utilize several objects prototypes, our strategy for setting d^f_0 is to minimize the false alarms for each prototype, and eliminate the missing objects using further feature tuples. By setting the F^{m}MPP *data* model parameters (d_0 and d_ψ), we utilized that using similar ISAR imaging conditions, the contrast parameters of the images are very similar. Finally, to set the *optimization* parameters, we followed the guidelines provided in [87] and used δ_0 between 10000 and 20000, β_0 between 20 and 50, and geometric cooling factors $1/0.96$.

4.5 Conclusions of the chapter

In this chapter, we proposed two different solutions for dynamic Marked Point Processes.

In the first part, we have proposed a multitemporal Marked Point Process (mMPP) framework for building extraction and change monitoring in remotely sensed image pairs in a joint probabilistic approach. A global optimization process attempted to find the optimal configuration of buildings, considering the observed data, prior knowledge, and interactions between the neighboring building parts. The computational cost has been significantly decreased by a non-uniform stochastic object birth process, which proposed relevant objects with higher probability based on low-level image features.

The second part has addressed the detection and characterization of large ship and airplane targets in ISAR image sequences using an energy minimization approach. We have proposed a robust joint model for axis extraction, feature point detection and tracking. We have shown that in case of noisy sequences, the introduced Multiframe Marked Point Process schema can significantly improve the results of frame-by-frame detection.

Chapter 5

Multi-level object population analysis with an Embedded MPP model

In this chapter we introduce a probabilistic approach for extracting complex hierarchical object structures from digital images used by various vision applications. The proposed framework extends conventional Marked Point Process (MPP) models by (i) admitting object-subobject ensembles in parent-child relationships and (ii) allowing corresponding objects to form coherent object groups, by a Bayesian segmentation of the population. Differently from earlier, highly domain specific attempts on MPP generalization, the proposed model is defined at an abstract level, providing clear interfaces for the possible applications. We also introduce a global optimization process for the multi-layer framework, which attempts to find the optimal configuration of entities, considering the observed data, prior knowledge, and interactions between the neighboring and the hierarchically related objects. The proposed method is demonstrated in three different application areas: built in area analysis in remotely sensed images, traffic monitoring on airborne Lidar data and optical circuit inspection.

5.1 A hierarchical MPP approach

In the recent years, one of the main evolving characteristic features of commercial perception sensors has been the spatial resolution. In the remote sensing domain, several very high resolution satellites have been launched including the Pleiades system (in 2011) which provides submetric resolution data incorporating stereo facilities. On the other hand, Full High Definition (HD) video cameras are available at affordable prices for many surveillance applications, which systems can be supported by arrays of multiple thermal or Time-of-Flight (ToF) sensors. The spatial resolution of airborne or terrestrial laser scanning is also increasing due to a focused development of aerial Lidar devices and mobile mapping systems. In industrial optical inspection systems research works on designing proper sources of illumination, decreasing lens aberrations and improving the limited depth of field result in sharp images up to a few μm resolution.

From the processing side, the above hardware developments imply a notable shift in computer vision methodologies. While several earlier technologies focused an compensating low image resolution e.g. by mosaicking or superresolution techniques, nowadays fine details are observable in high-resolution images, demanding hierarchical content parsing algorithms [72], which can interpret the observed information at multiple levels. While the conventional MPP-based image analysis models (see Sec. 2.2 and also our earlier introduced multitemporal models in Chapter 4) are purely focusing on object extraction with direct (bilateral) object interaction modeling within populations, dealing with higher level object grouping and object decomposition issues are also critical parts of complex scene understanding processes.

There have already been a few attempts conducted for multi-entity-level modeling with point processes in the literature. The Multi-MPP framework proposed by [109] offers extensions of MPP models regarding two issues. *First*, to simultaneously detect variously shaped entities, it jointly samples different types of geometric objects. *Second*, by a statistical type and alignment analysis of the extracted nearby entities local texture representation of the different image regions is obtained. Although this approach fits well to bottom-up exploration tasks of the unknown imaged scene content, it is not straightforward in many vision applications, how to efficiently segment in this framework the object population based on domain specific top-down knowledge. On the other hand, several hierarchical phenomena can be better described by object-subobject ensembles in parent-child relationships rather than by object grouping constraints. As examples, we can mention here Circuit Elements (CE) of Printed Circuit Boards (PCB) and recognizable patterns of included artifacts within the CEs [8, 11] in μm resolution images, building roofs and chimneys in aerial or satellite photos, ships and containers in radar images [7] etc.

For the above reasons, we introduce in this chapter a new three level Embedded Marked Point Process (EMPP) framework [2, 25], which has the following two key properties:

- We describe the hierarchy between objects and object parts as a parent-child relationship embedded into the MPP framework. The appearance of a child object is affected by its parent entity, considering geometrical and spectral constraints, such as the geometric figure of a parent object encapsulates figure of its a child object, or the color/texture of the parent object may influences the appearance characteristic of the child entity.

- For avoiding the limitations of using pairwise object interactions only, we propose here a multi-level MPP model, which partitions the complete (parent) entity population into object groups, called configuration segments, and extracts the objects and the optimal segments simultaneously by a joint energy minimization process. Object interactions are differently defined within the same segment and between two different segments, implementing adaptive object neighborhoods. In this way, we can use in parallel strong alignment or spectral similarity constraints within a group, but the coherent segments may even have irregular, or thin, elongated shapes.

As it will be shown in the following sections, the EMPP model has a complex structure, with several general and task specific components mixed together in a unified framework. Practical experiences show that for such composite, application dependent models, the adaption to another application domain is rarely straightforward, and usually a significant amount of modeling work and code (re-)implementation is needed to transform or modify the framework for a different field. As an important novelty of our present method, after collecting similar connecting tasks appearing in different areas, we address them by a joint methodological approach. We provide here a formal problem statement and introduce a novel general three-level MPP framework which enables us to handle a wide family of applications. The structure elements and the energy optimization algorithm of the complex model are defined and implemented at the abstract level, while we keep focus on ensuring very simple interfaces to the different applications, providing flexible options for domain adaption for end-users.

The development of the EMPP model contained three phases. *First*, we proposed an Automatic Optical Inspection (AOI) method for PCB validation, with introducing the parent-child relationship into the conventional MPP framework. *Second* we designed a two-level MPP model focusing on the joint extraction of vehicles and groups of corresponding vehicles within a traffic scenario from aerial Lidar data. Both models have been thoroughly validated on real measurements, and the methodological improvements have been demonstrated versus earlier approaches from the literature. In the *third* phase, we have connected the two models, and defined the general three layer EMPP framework containing the object group - object - object part levels. Thereafter we gave proof-of-concept examples how the new three layer model can be adopted to the targeted AOI, traffic monitoring and built-in area estimation applications. The Author's scientific contributions

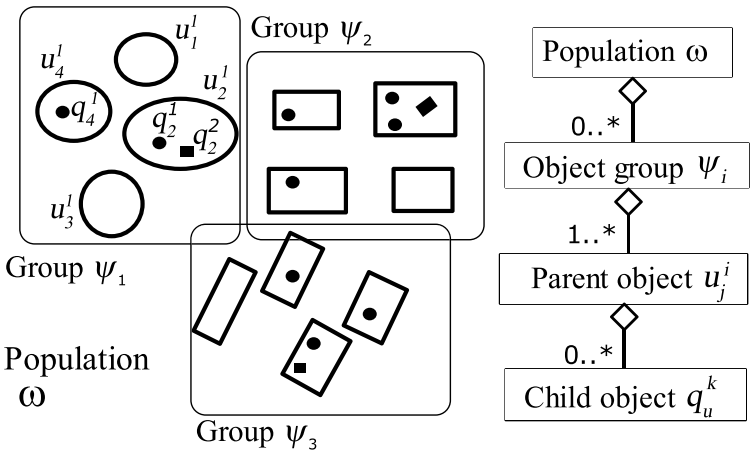


Figure 5.1: Structure elements of the EMPP model. Left: a sample population with three object groups, and various object shapes both at parent and child layers. Right: The multi-layer structure of the model featuring the encapsulation relation.

regarding the first step are formulated separately from the later two ones in the sub-theses of Thesis 3 (see Chapter 7). However, for a more compact and less redundant presentation, we take a reverse order in this Chapter, starting with the introduction of the general EMPP model, an we detail the application specific contributions regarding the AOI and traffic monitoring tasks thereafter.

5.2 Problem formulation and notations

To model the hierarchical scene content, the proposed Embedded Marked Point Process (EMPP) framework has a multi-layer structure, as shown in Fig. 5.1. At the top, we have a super node, called the *population* or the *configuration*, which is a high-level model of the imaged scene. The population consists of an arbitrary number of object groups, where each group is a composition of one or many super (or parent) objects. Finally, the super objects may encapsulate any number of subobjects (or child objects).

Following the notations introduced in Chapters 2 and 4, the input of the EMPP method is an image over a pixel lattice S, and $s \in S$ denotes a single pixel. As the first extension of conventional MPP models, each parent object $u \in \mathcal{H}$ may contain a set of child objects $Q_u = \{q_u^1 \dots q_u^{m(u)}\}$ where $m(u) \leq m_{\max}$ and $q_u^i \in \mathcal{H}$. $Q_u = \emptyset$ marks that u has no child. Let us denote by $\mathcal{H}_Q$ the parameter space of all possible Q_u vectors. Both the parent and child objects are represented by plane figures from preliminary defined shape libraries.

As for the second level of the proposed object hierarchy, we introduce the object grouping process. A given population, denoted by ω, is a set of k *object groups* or (also referred later as

configuration segments), $\omega = \{\psi_1, \ldots, \psi_k\}$, where each group ψ_i ($i = 1 \ldots k$) is a configuration of n_i objects:

$$\psi_i = \{u_1^i, \ldots, u_{n_i}^i\} \in (\mathcal{H} \times \mathcal{H}_Q)^{n_i}. \tag{5.1}$$

Here we prescribe that $\psi_i \cap \psi_j = \emptyset$ for $i \neq j$, while the k set number and $n_1, \ldots, n_k$ set cardinality values may be arbitrary (and initially unknown) integers. We mark with $u \prec \omega$ if u belongs to any ψ in ω, i.e. $\exists \psi_i \in \omega : u \in \psi_i$. Let us denote by $\mathcal{N}_u(\omega)$ the proximity based neighborhood of $u \prec \omega$, which is independent of the group level: $\mathcal{N}_u(\omega) = \{v \prec \omega : u \sim v\}$.

Finally, we denote by Ω the space of all the possible global configurations, which is constructed as:

$$\begin{gathered}\Omega = \cup_{k=0}^{\infty} \left\{ \{\psi_1, \ldots, \psi_k\} \in [\cup_{n=1}^{\infty} \Psi_n]^k \right\} \\ \text{where } \Psi_n = \left\{ \{u_1, \ldots, u_n\} \in (\mathcal{H} \times \mathcal{H}_Q)^n \right\}.\end{gathered} \tag{5.2}$$

In this way, we consider that each population $\omega \in \Omega$ may include any number of groups composed of any number of objects and child objects.

5.3 EMPP energy model

The EMPP energy function is derived with minor modifications of the basic formula (2.21):

$$\Phi(\omega) = \Phi_d(\omega) + \gamma \cdot \Phi_p(\omega) \tag{5.3}$$

where $\Phi_d(\omega) = \sum_{u \in \omega} A(u)$ is the unary term and $\Phi_p(\omega)$ is the prior interaction term.

The *unary term* construction process follows the same approach as introduced in Sec. 4.2.2.2, with the only difference that in the EMPP model, the $A(u)$ data energy function is decomposed into a parent term $\varphi_d^p(u)$ and child terms $\varphi_d^c(u, q_u)$. As indicated by the notation, a child term may depend both the local image data and the geometry of the parent object (e.g. an intensity histogram within the parent region). Both the parent and the child level energy components are derived according to the earlier introduced schema: *First* fitness features are derived to characterize the efficiency of the generated super/sup-object candidates respectively. *Second*, the features are mapped with the nonlinear $\mathcal{M}(f, d_0, D)$ function to obtain the energy subterms corresponding to the f feature. *Third* the joint data energy of object u is derived by combining averaging, max and min operators, using the multiple prototype definition strategies presented in Sec. 4.2.2.2

The complete unary term of u is the sum of the parent level terms and the child level terms:

$$A(u) = \varphi_d^p(u) + \sum_{q_u \in Q_u} \varphi_d^c(u, q_u) \tag{5.4}$$

The *interaction terms* implement geometric or feature based interaction constraints between different objects, child objects and object groups of ω.

$$\Phi_p(\omega) = \underbrace{\sum_{u \sim v} I_p(u,v)}_{\text{parent-parent interaction}} + \underbrace{\sum_{u \prec \omega} I_c(u, Q_u)}_{\text{parent-child interaction}} + \underbrace{\sum_{u,\psi} I_g(u,\psi)}_{\text{parent-group interaction}} \tag{5.5}$$

First, the $I_p(u,v)$ terms provide classical pairwise interaction constraints, for example, we can use the common intersection term from eq. (2.22), which penalizes overlapping objects: $I_p(u,v) = \frac{\#\{R_u \cap R_v\}}{\#\{R_u \cup R_v\}}$.

Second, the $I_c(u, Q_u)$ terms model interactions between the corresponding parent and child objects, and interactions between different child objects corresponding to the same parent. For example, we can prescribe that the children of a given parent (i.e. *siblings*) should not overlap with each other, and not overhang the parent, or the siblings should have same shape, similar color, size, orientation etc.

Third, with the $I_g(u,\psi)$ energies, can define various constraints between the object group level and the (parent) object level of the scene [28]. To measure if an object u appropriately matches to a population segment ψ, we define a distance measure $d_\psi(u) \in [0,1]$, where $d_\psi(u) = 0$ corresponds to a high quality match. In general, we prescribe that the segments are spatially connected, therefore, we use a constant high difference factor, if u has no neighbors within ψ w.r.t. relation $\sim$. Thus we derive a modified distance:

$$\hat{d}_\psi(u) = \begin{cases} 1 & \text{if } \nexists v \in \psi \backslash \{u\} : u \sim v \\ d_\psi(u) & \text{otherwise} \end{cases} \tag{5.6}$$

By definition of $I_g(u,\psi)$, we slightly penalize population segments which only contain a single object:

$$I_g(u,\psi) = c \text{ iff } \psi = \{u\}, \tag{5.7}$$

with a small $0 < c$ constant.

For segments with multiple objects, we penalize large $\hat{d}_\psi(u)$ distances within a group, and also small $\hat{d}_\psi(u)$ distances if u is not a member of ψ:

$$I_g(u,\psi) = \begin{cases} \hat{d}_\psi(u) & \text{if } u \in \psi \\ 1 - \hat{d}_\psi(u) & \text{if } u \notin \psi. \end{cases} \tag{5.8}$$

5.4 Multi-level MPP optimization

For optimizing the energy function of eq. (5.3), we have have extended again the Multiple Birth and Death (MBD) [87] algorithm. To accommodate to the requirements of the EMPP energy function, the main task was to include the group assignment, object re-grouping, and child maintenance

issues within the original MBD framework. On one hand, after each *birth* step, the generated object should be assigned to a new, or an existing group. Then, following the *death* procedure, we execute a new step, called *Group re-arrangement*, which may re-direct some objects to neighboring object groups based on data based and prior soft-constraints. On the other hand, in the last step of an iteration, called *Child Maintenance*, we may add, remove or replace child objects for each parent. As already discussed in Chapter 4, efficient object proposal strategies can significantly speed up the MPP energy optimization algorithms. While the *Feature Based Birth Process* (FBB) introduced in Sec. 4.2.4 proved to be efficient for the building detection application, we have also introduced an extended schema, called the *Bottom-Up Stochastic Entity Proposal* (BUSEP) process, which has been successfully adopted first to Printed Circuit Board (PCB) inspection [8], then to Lidar-based vehicle detection applications. The BUSEP algorithm is executed as a preprocessing step of the iterative optimization, where we assign to the different image pixels (1) pseudo probability values that the pixel is an object reference point (e.g. center of an ellipse) (2) narrow distributions for all object parameters (including orientation and side/axis length parameters) expected in the given pixels, based on a deterministic *object candidate extraction* procedure. During the generation of a new object in the birth step, we follow the distributions of the expected parameters, which yields that efficient candidates will be proposed significantly faster. On the other hand, similarly to the birth maps [87] and FBB strategies, the entity proposal maintains the reversibility of the iterative evolution process of the object population [110], instead of implementing a greedy algorithm. We use as input of BUSEP a binary *foreground* mask obtained by a task specific deterministic segmentation algorithm from the input image, which realizes a coarse separation of the parent or child objects from the background. Technical details of the proposed EMPP optimization process are provided in Appendix D: an example for efficient candidate generation for BUSEP in the optical PCB inspection application is presented in Algorithms D.1-D.4, while the pseudo code of the proposed new Multi-level Multiple Birth-Death-Maintenance ($\mathrm{M^M}$BDM) algorithm is shown in Algorithm D.4.

5.5 Applications of the EMPP model

Implementing the interfaces of the EMPP framework consists of specifying the following issues for each application:

Model elements: semantic definition of parent/child objects and object groups. Fixing the shape libraries for parent/child objects, and additional domain specific constraints such as the maximum number of *siblings* having the same parent.

Unary terms: defining the domain specific f features and feature integration rules to obtain the *parent level* $\varphi_Y^p(u)$ and *child level* $\varphi_Y^c(u, q_u)$ unary terms.

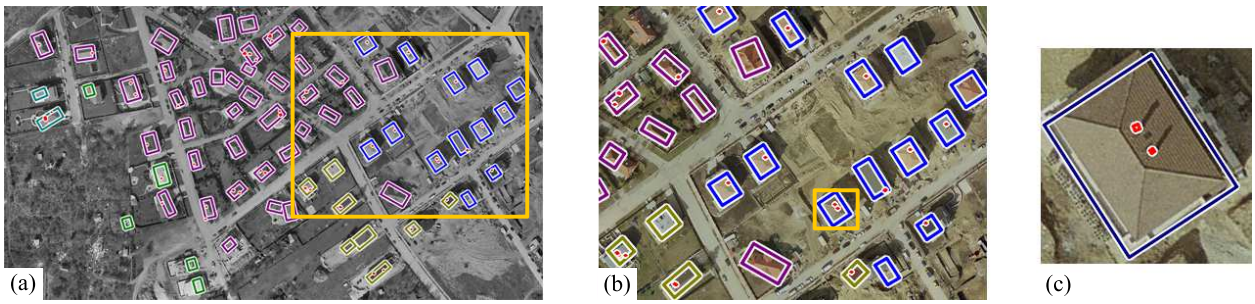


Figure 5.2: Results of built-in area analysis, displayed at three different scales. Building groups are distinguished with different colors (purple: *red roofs' district*, others: orientation based groups); red markers denote the detected chimneys

Parent-parent interactions: defining the $I_p(u, v)$ interaction terms between (spatially) neighboring parent objects.

Parent-child interactions: defining the $I_c(u, Q_u)$ interaction constraints between the corresponding parent a children objects.

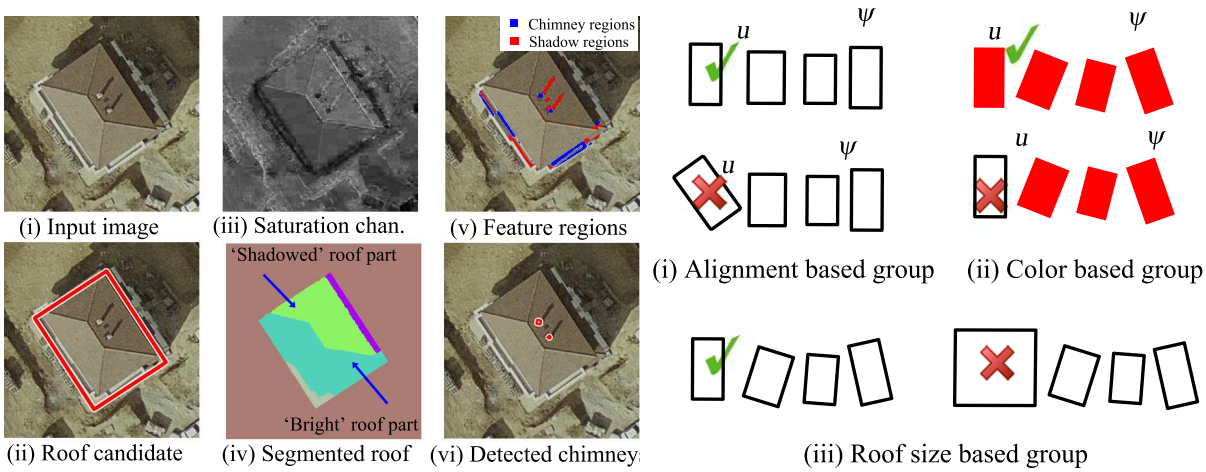
Parent-group interactions: defining the grouping constraints through the definition of the $d_\psi(u)$ object-segment distance.

We emphasize hereby that all further model elements and algorithmic steps introduced in Sections 5.2-5.4 are independent on the concrete application, which property was ensured during model implementation by a clear separation of the general and tasks specific program components.

5.5.1 Built-in area analysis in aerial and satellite images

As we introduced in Chapter 4, analyzing built-in areas in aerial and satellite images is a key issue in several remote sensing applications, e.g. in cartography, GIS data management and updating, or disaster management. Most existing techniques focus on the extraction of individual buildings or building segments from the images [10], however, as pointed out in [150] finding groups of corresponding buildings (e.g. a residential housing district) has also a great interest in urban environment planning, as well as detecting illegally built objects which do not fit the regular environment. On the other hand authorities or telecommunication companies may also need to monitor specific objects on the roofs such as chimneys or parabolic antenna dishes for either statistical purposes (market research), or for the estimation of air pollution. Detecting illegal or irregular chimneys can also be a relevant task for city monitoring.

For demonstrating the adaptation of the EMPP model for the topic of urban area analysis, we have chosen very high resolution aerial images (around 12cm/pixel) captured from regions of Budapest, Hungary, a representative sample is displayed in Fig. 5.2. The task specific issues are detailed in the following.

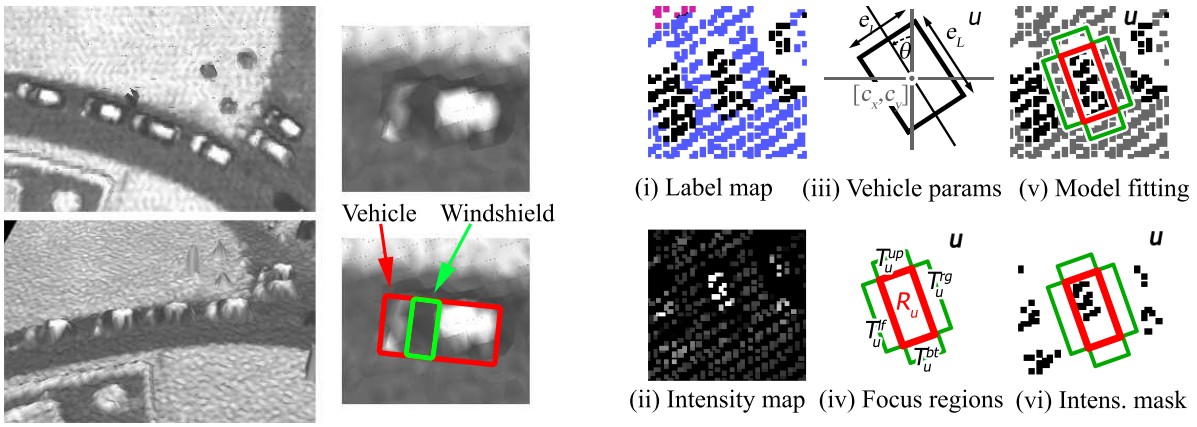


(a) Features for chimney extraction (b) Types of used building groups

Figure 5.3: Built-in area analysis - model components

Model elements in built-in area monitoring: Parent objects are rectangular segments of the building footprints, assuming that each building can be approximated from the top-view either by a rectangle or by a couple of slightly overlapping rectangles. Child objects are tall structure elements on the roofs, such as chimneys or satellite dishes, also modeled by rectangles. For easier discussion, we refer to all child objects simply as *chimneys* in the following. Configuration segments are groups of corresponding buildings, like members of a residential housing district in Fig. 5.2(a).

Unary terms of buildings and chimneys: Parent level unary terms are derived in the same way as introduced in Chapter 4: the energy function integrates feature-information about roof color, roof edge and shadow. As for the *child unary terms*, the feature extraction workflow for indicating chimneys (or further tall structure elements) on the roofs is demonstrated in Fig. 5.3(a). We used two observations. First, chimney pixel colors have usually lower saturation components compared to the surrounding roof parts, which can be filtered in the HSV color space considering the *saturation channel* (Fig. 5.3(a)-(iii)). Second, chimneys cast shadows on the roofs, an issue which can be approached in a manner similar to localizing buildings using the shadows on the parent object level. However, for non-flat roofs (such as gable or mansard roofs [82]) we must separately handle the cases of illuminated and self-shadowed roof segments. Taking a photometric approach [14], for a given surface point the ratio of the observed intensities (luminance or gray level) in shadow and under illumination may be efficiently modeled by a Gaussian density function in outdoor scenes. However, the mean value of the Gaussian varies according to external illumination [14], i.e. it needs different settings for the illuminated and shadowed roof parts. Thus, we first segment the parent object region using a floodfill-based classification step (Fig. 5.3(a)-(ii)(iv)), then



(a) Vehicle by Lidar (b) Unary term features

Figure 5.4: Vehicle detection from airborne Lidar data: (a) vehicles appearances in raw triangulated Lidar data (intensity based coloring was used), (b) calculation of the data model features

a local color model is adopted in each segment, derived from the regions' histograms. The estimated chimney object and shadow regions are shown in Fig. 5.3(a)-(v) with blue and red overlays, respectively. Finally the child object's data term prescribes *chimney candidate* pixels within the object mask and *shadowed* areas in the neighboring roof regions w.r.t. the global shadow direction. Examples for extracted chimney objects are shown in Fig. 5.3(a) and Fig. D.2.

Parent-child terms $J(u, Q_u)$: Non-overlapping siblings are expected to have similar orientation. Children figures should be encapsulated by the parent rectangles (Fig. 5.2(c)).

Object-segment distance $\hat{d}_\psi(u)$: In our test areas, we have observed various different grouping constraints, which should be considered on a case-by-case basis. First, in many regions, we can find several distinct building groups which are formed by regularly aligned, parallel buildings. Second, we can also see large building groups (e.g. purple group in the center of Fig. 5.2(a)), where the orientations of the houses are irregular, but the roof colors are uniform. Third, family houses and condominiums can be mixed in the same area, which can also be a basis for grouping. Thus, we distinguished three types of building groups: if ψ is an alignment based group (Fig. 5.3(b)-(i)), $d_\psi(u)$ is proportional to the angle difference between u and the mean angle within ψ. Otherwise, if ψ is a color group (Fig. 5.3(b)-(ii)), $d_\psi(u)$ measures how the color histogram of u matches the ψ group's expected color distribution, which is set by training samples during the system configuration. Finally, for separating individual houses from larger condominiums, the roof size and the side length ratios are the discriminative features (Fig. 5.3(b)-(iii)).

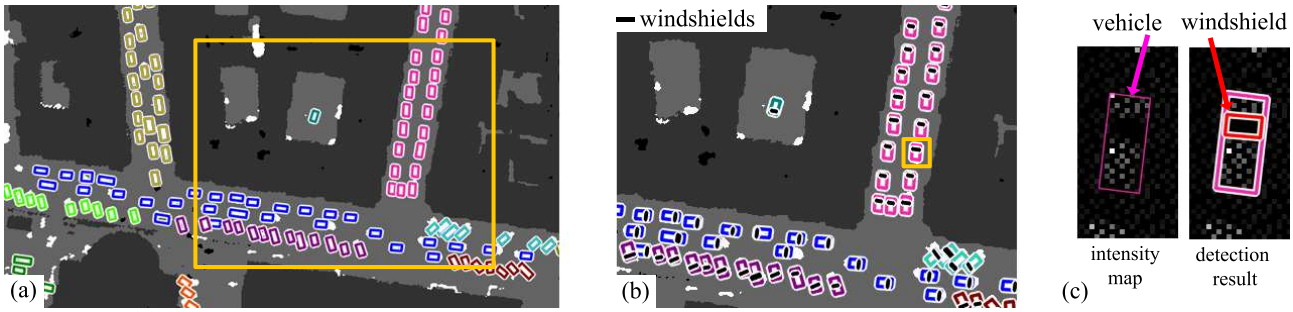
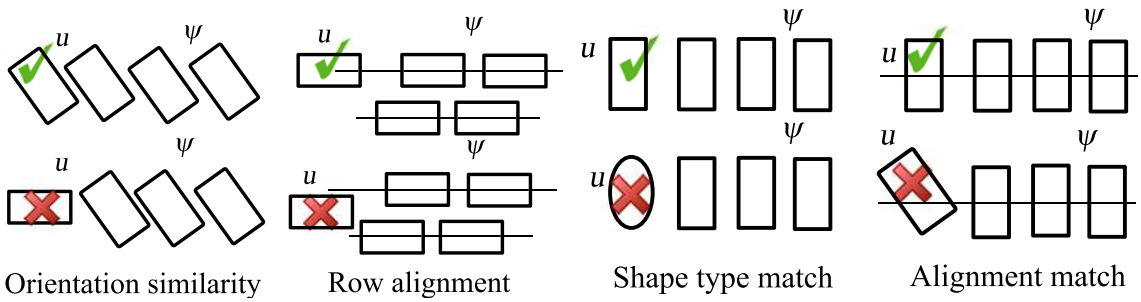


Figure 5.5: Sample results on traffic analysis. Super rectangles mark the detected vehicles, different colors correspond to the different groups. In the background, gray levels refer to the input label map: white - vehicle candidates, light gray - road, dark gray - roof. a) cars and traffic segments b) selected region with the detected windshields c) intensity map of a selected car, d) detection result for c).

5.5.2 Traffic monitoring based on Lidar data

In city surveillance applications, automatic traffic monitoring and analysis needs a hierarchical modeling approach: first *individual vehicles* should be detected, then we need to extract *coherent traffic segments*, by identifying groups of corresponding vehicles, such as cars in a parking lot, or a vehicle queue waiting in front of a traffic light. In addition, extracting characteristic parts of the vehicles may provide useful information for classification or behavior analysis. In this section, we rely on the measurements of an airborne Lidar laser scanner and a car-mounted mobile mapping system (MLS), providing 3D point clouds completed with intensity/RGB color values. From the aerial data, due to the low resolution of the considered point cloud measurements (max. 8 points/m^2), only coarse vehicle shapes can be extracted. However, as shown in Fig. 5.4(a), the windshields are observable, so they could be separated based on a joint consideration of the vehicle geometry and the observed intensity map. From a practical point of view, extracted windshields can be used for classifying vehicle types, estimating vehicle direction etc. As for the MLS data (Fig. 5.7), the point cloud has a very high resolution, preserving several details, but significant challenges are caused by ghost objects, occlusion and invisible object parts, which are the consequences of the street level scanning process.

In [5] we introduced a two-step method for Lidar based vehicle detection, which is adapted and extended here for the EMPP framework. Firstly, each point of the 3D point set is classified into vehicle or background clusters, however, this classification can only be considered as a coarse input for the object detector. Then the points with the corresponding class labels and intensity values are projected to the ground plane, where the optimal vehicle and traffic segment population is modeled by a rectangle configuration in the projected 2D image. A sample class label map extracted from aerial data is demonstrated in Fig. 5.5(a), while the projected intensity map of an MLS data segment is shown in Fig. 5.7(c).



(a) Vehicle groups (b) PCB groups

Figure 5.6: Grouping energies for (a) traffic monitoring and (b) printed circuit analysis applications. Favored ($\surd$) and penalized ($\times$) sub-configurations within an object group

Model elements: parent objects are vehicles, child objects are windshields (both are rectangles). Configuration segments are formed by corresponding vehicles according to various traffic situations (Fig. 5.5(a)).

Parent unary terms (φ_Y^p): as we introduced in [5], three different features are exploited for vehicle extraction (see Fig. 5.4(b)). The *vehicle evidence* (f_{ve}) respectively *intensity* (f_{it}) features are calculated as the covering ratios of vehicle classified pixels in the label and intensity maps within the proposed rectangle of u. The *external background* (f_{eb}) feature is the rate of background classified pixels in neighboring regions around the proposed u object. The ϕ_{ve}, ϕ_{it} and ϕ_{eb} primitive terms are derived according to eq. (4.1), similarly to the built-in area analysis application. Finally the joint data energy of object u is calculated as:

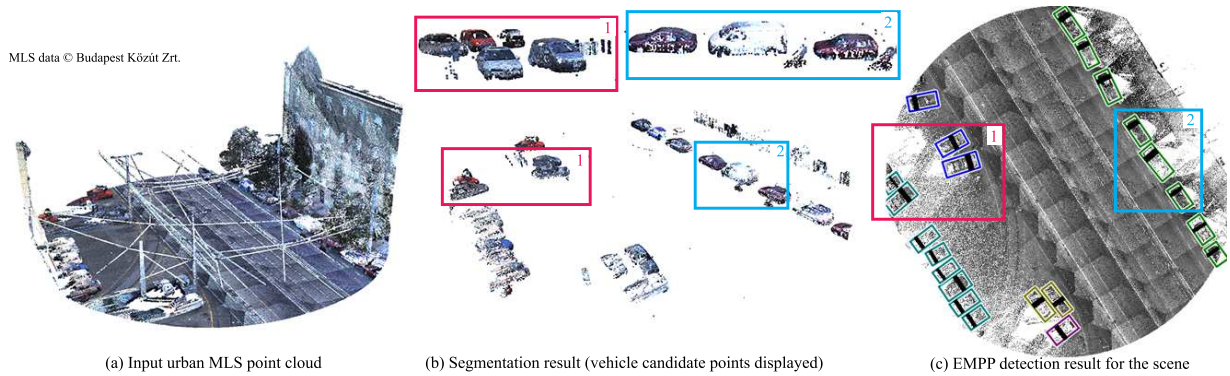
$$\varphi_Y^p(u) = \max(\min(\phi_{\text{it}}(u), \phi_{\text{ve}}(u)), \phi_{\text{eb}}(u)), \tag{5.9}$$

where we admit that not necessarily all vehicles appear as bright blobs in the intensity map.

Child unary terms (φ_Y^c): due to their glassy material, the windshield rectangles cover regions without points or low-intensity areas in the projected point cloud maps (Fig. 5.4(a) and 5.5(c)), features which are characterized by coverage ratios similarly to the parent level descriptors.

Parent-child terms $J(u, Q_u)$: the windshield is encapsulated by the car's figure, and the orientation is perpendicular to the car's main axis (Fig. 5.5(c)).

Object-segment distance $d_\psi(u)$: we expect that the vehicles of the same segment have similar orientations, and they form regular rows. The $d_\psi(u)$ distance is the average of two terms: the *first* term is the normalized angle difference between u and the mean angle within ψ (see Fig. 5.6(a)-left). Regarding the *second* term, we fit one or a couple of parallel lines to the object centers within ψ using RANSAC, and calculate the normalized distance of the center of u from the closest line (Fig. 5.6(a)-right). A generalization of this feature for curved road segments can be found in [5].



(a) Input urban MLS point cloud (b) Segmentation result (vehicle candidate points displayed) (c) EMPP detection result for the scene

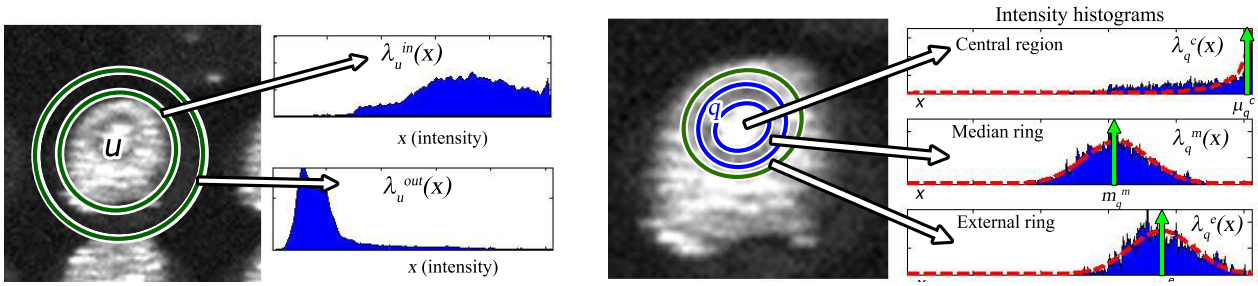
Figure 5.7: Processing workflow for Mobile Laser Scanning (MLS) data. (a) Input scene (b) estimated vehicle regions by point cloud classification - two selected segments are highlighted from different viewpoints (c) EMPP detection results

5.5.3 Automatic optical inspection of printed circuit boards

Automatic optical inspection (AOI) is a widely used approach for quality assessment of Printed Circuit Boards (PCBs). Automated layout-template-free approaches are especially useful for verifying uniquely designed circuits. In the PCBs usually connected groups of similarly shaped and oriented Circuit Elements (CEs) implement a given function, therefore the interpretation of the board content needs to segment the CE population. Another critical issue is filtering the flawed PCBs by AOI. Nowadays the most widespread assembling technology of electronic circuit modules uses reflow soldering [151]. Here a common problem, called *scooping* may occur during manufacturing, which influences the strength of solder joints in stencil prints [8]: a board should be withdrawn if the number the summed volume of such artifacts surpass a given threshold. A scoop can be visually observed in an AOI image as a bright patch surrounded by a darker ring within the solder paste, as shown in Fig. 5.8(a). Automatic detection is challenging due to the locally varying contrast of AOI images. In [11], I proposed an initial Bayesian approach for joint extraction of the circuit elements and the included scoop. Afterward, in [8] we presented a deep study about the technological background of this artifact, and proposed an advanced solution, called the *Hierarchical Multi Marked Point Process* (H^MMPP) method, to cope with the complex PCB analysis task. In this thesis we briefly demonstrate only, how our scooping detection approach can be adapted to the EMPP framework, but we recommend the Reader to also study our related publications [8, 11, 37].

Model elements: parent objects are CEs of various shapes, child objects are scoops, modeled by pairs of concentric ellipses. Groups are formed by CEs which likely have similar functionalities.

Parent unary terms (φ_Y^p): In the considered PCB image data set [11] the CEs can be modeled as bright *rectangles*, *ellipses* or *triangles* surrounded by darker background. To evaluate the contrast between the CEs and the board, we calculate the Bhattacharya [87] distance $d_B(u)$ between the



(a) Parent level features (b) Child level features

Figure 5.8: PCB inspection: Feature demonstration for unary term calculation

pixel intensity distributions of the internal CE regions and their boundaries (see Fig. 5.8(a)). Then the $\varphi_Y^p(u)$ unary term is derived by $\mathcal{M}$ mapping of $d_B(u)$.

Child unary terms (φ_Y^c): Following the approach of [8] we distinguish three regions of each scoop: the central bright ellipse, the darker median ring and the bright external ring, as shown in Fig. 5.8(b). Experimental evidences prove here, that for a real scoop q, the gray level histogram of the central region, $\lambda_q^{\mathrm{c}}(x)$ follows a skewed distribution, while the medium and external region histograms ($\lambda_q^{\mathrm{m}}(x)$ resp. $\lambda_q^{\mathrm{e}}(x)$) can be approximated by Gaussian densities. Let us denote by μ_q^{c}, μ_q^{m} resp. μ_q^{e} the peak locations of the smoothed $\lambda_q^{\mathrm{c}}(x)$, $\lambda_q^{\mathrm{m}}(x)$ resp. $\lambda_q^{\mathrm{e}}(x)$ functions. We prescribe three constraints for an efficient scoop candidate: (i) it exhibits high μ_q^{c} value; while intensity ratios (ii) $\mu_{q_u}^{\mathrm{c}}/\mu_{q_u}^{\mathrm{m}}$ resp. (iii) $\mu_{q_u}^{\mathrm{e}}/\mu_{q_u}^{\mathrm{m}}$ pass given contrast thresholds d^{cm} and d^{em}. To enforce the simultaneous fulfillment of the (i)-(iii) properties, the child's data-energy value is calculated applying the maximum operator (logical AND) from the subterms of the three constraints. We use here again the $\mathcal{M}$ function, defined by eq. (4.1):

$$\varphi_Y^c(u, q_u) = \max\left(\mathcal{M}(\mu_{q_u}^{\mathrm{c}}, d^c), \mathcal{M}(\mu_{q_u}^{\mathrm{c}}/\mu_{q_u}^{\mathrm{m}}, d^{\mathrm{cm}}), \mathcal{M}(\mu_{q_u}^{\mathrm{e}}/\mu_{q_u}^{\mathrm{m}}, d^{\mathrm{em}})\right) \tag{5.10}$$

Parent-child terms $J(u, Q_u)$: due to the manufacturing technology at most one scoop may appear in a solder paste, therefore each parent CE may have a maximum of one child, whose figure cannot overhang its parent.

Object-segment distance $d_\psi(u)$: within a CE group, we prescribe that the elements must have similar shape and must follow a strongly regular alignment (Fig. 5.6(b)). Therefore $d_\psi(u) = 1$ if the type of u, $\mathrm{tp}(u)$ is not equal to the type of the ψ group, otherwise $d_\psi(u)$ is the maximum of the angle difference and symmetry distance terms defined in Sec. 5.5.2 by the traffic monitoring application.

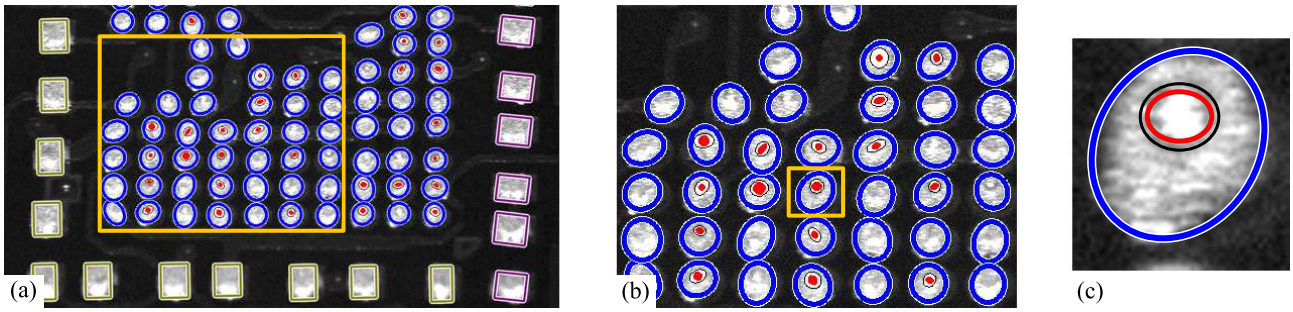


Figure 5.9: Results of PCB analysis. CEs are grouped by shape and orientation, scoops are extracted within the CEs

5.6 Benchmark database and evaluation methodology

Since to our best knowledge no usable dataset has been published yet enabling the three-level analysis of the discussed complex scenarios, we have created the new *EMPP Benchmark database*[1], which is designed for the evaluation of multilevel object population analysis techniques on high resolution images. For each image Ground Truth (GT) data has been generated, which encodes the dependencies of objects, object groups and child objects within a population. For GT annotation we have developed a program with graphical user interface, which enables the user to manually create and edit a GT configuration of various geometric objects composed of both parent and child elements. We can also create new object groups, and assign each parent object to an existing group.

The EMPP Benchmark database includes the following input images with annotation:

Building detection: Budapest aerial image with 12cm resolution (69 buildings, 79 chimneys), Manchester satellite image (50cm res., 155 buildings) from the SZTAKI-INRIA Benchmark (see Sec. 4.2.4, [10]), and two Quickbird images (#2 and #11, 60cm-80cm res., 218 buildings) from the dataset by A.O. Ok [152].

Traffic analysis: the dataset contains aerial Lidar point clouds, and some mobile laser scanning (MLS) data samples from dense urban regions of Budapest, Hungary. The *aerial data* part consists of 6 point cloud segments including 792 vehicles [5], while the *MLS data* includes 2 point cloud segments with 42 vehicles (scanner: Riegl VMX-450 mobile mapping system).

Optical circuit board analysis: a large dataset of 44 printed circuit board images with 6μm resolution, containing 4439 CEs and 664 scooping errors [8].

The quantitative evaluation of an EMPP based scene analysis algorithm should be accomplished at multiple levels. In the *parent object* layer, we use both object based and pixel based accuracy rates in the same way as defined in Section 4.2.4. The evaluation step regarding the the *child layer* uses object level metrics similarly to the parent layer. However, by calculating the Child

[1]Url: `http://mplab.sztaki.hu/EMPPBenchmark`

Table 5.1: Object and group level evaluation of the the proposed EMPP model, and comparison to a conventional *s*MPP approach

App.	Method	Parent level analysis						Group level study	
		Number of objects			Pixel level %			Obj mis-grouping	
		TP	FP	FN	PRc	PPr	**PPF**	FG#	GR%
Building analysis	*s*MPP	406	24	36	80	75	**78**	58	14
	EMPP	417	14	25	84	88	**86**	28	7
Aerial traffic monitoring	*s*MPP	792	30	25	79	77	**78**	202	25
	EMPP	793	30	24	82	85	**83**	43	5
Ground-based traffic analysis	*s*MPP	42	0	0	92	86	**89**	2	5
	EMPP	42	0	0	96	89	**92**	0	0
PCB inspection	*s*MPP	4408	39	31	87	86	**87**	448	10
	EMPP	4415	9	24	92	97	**94**	137	3

Object level Precision (CPr), Recall (CRc) and F-score (COF), we only accept matches between the detected and GT child objects, if their parents are also correctly matched at the upper layer. Finally, we also measure the correct Group Classification Rate (GR, %) among the true positive samples, considering the GT group classification information. The GR value is determined by counting the number correctly grouped objects (TG), the number of falsely grouped objects (FG), and calculating GR=TG/(TG+FG).

5.7 Experimental results

We evaluated our method on the new EMPP Benchmark database. Qualitative sample results of the three level population detection are shown in Fig. 5.2, 5.5, 5.7 and 5.9. During the quantitative analysis, the results were compared to the GT configuration of the benchmark, and the above performance rates were calculated in each case, as shown in Table 5.1 and 5.2.

During the first part of the comparative tests, we focused on the evaluation of the newly introduced EMPP framework versus earlier straightforward MPP solutions. As a baseline for comparison, we implemented a sequential technique, which extracts first the object population by a *s*ingle layer MPP model (*s*MPP), using exactly the same unary terms and child detection process as the proposed EMPP approach, but the $\Phi_p(\omega)$ prior term is only composed of the $I(u, v)$ intersection component and the $J(u, Q_u)$ parent-child interaction feature, while the parent-group term is considered to be zero ($A(u, \psi) = 0$). Thereafter, the parent object grouping step is performed in post processing by a recursive floodfill-like segmentation of the population. Starting from a randomly chosen object, we assign all its spatial neighbors to the same cluster iff the difference between the

Table 5.2: Child level evaluation of the the proposed EMPP model

Application	CRc%	CPr%	**COF**%
Building analysis	80	71	75
Aerial traffic monitoring	92	92	92
Ground-based traffic analysis	93	93	93
PCB inspection	91	95	93

orientations is lower than a τ threshold and recursively repeat the process until all objects receive a group label. As observed during the following qualitative and quantitative tests, the bottleneck is the usage of this single τ threshold, which cannot be set uniformly for a complete population in case of noisy initial object estimations.

In Table 5.1 we can observe that the introduced EMPP model can surpass *s*MPP in two major quality factors. First, EMMP results in a notable gain in the pixel based error rates (PRc, PPr and PPF), which means that the extracted object shapes become more accurate. Second, the EMPP model significantly decreases the number of objects with False Groups (FG,GR). Using the single layer model the main source of errors is that in many cases the object orientations cannot be accurately estimated based on the input feature maps only: in the building analysis task the edge map is often weak and noisy, in aerial vehicle detection the projected point cloud has a low resolution, and in PCB analysis the irregular deformations of the rectangular solder pastes may make the estimation inaccurate. On the other hand, in our EMPP model, the object orientations are efficiently adjusted by considering the higher (group) level alignment constraints. As shown in Table 5.1, the differences between the *s*MPP and EMPP performance are less significant regarding the Mobile Laser Scanning (MLS) data, which has a high resolution and accuracy, enabling more reliable feature extraction from the input measurements. We note that in particular cases, the *s*MPP output could also be enhanced by using pairwise orientation smoothing terms [109]. However, the proposed EMPP model offers a higher degree of freedom for simultaneously considering various group level features and exploiting interaction between corresponding, but not necessarily closely located objects. In our case, we only prescribe regular alignment within the estimated object groups, locally outlying labels can indicate unusual object behavior.

Another relevant point of evaluation is the justification of using an MPP approach versus various alternative non-MPP based techniques for the selected application domains. Regarding the *building detection* problem, a detailed state-of-the comparison has already been provided in Chapter 4, which demonstrated the advantages of the point process based solution.

Vehicle detection from airborne Lidar has also a broad literature. In our task specific paper [5], we compared our solution to the digital elevation map based PCA [153], h-maxima suppression (h-max) [154] and Floodfill (FF) [32] approaches. Although the reference methods were chosen so

Table 5.3: Traffic analyis evalutation vs. state-of-the-art. Parent level F-scores (in %) by the PCA [153], h-max [154], Floodfill (Floodf) and the proposed EMPP methods.

Set	NV*	Object level F-score %				Pixel level F-score%			
		PCA	h-max	Flof	**EMPP**	PCA	h-max	Flof	**EMPP**
#1	191	78	78	88	**97**	63	63	66	**82**
#2	94	89	81	80	**97**	80	38	60	**73**
#3	170	85	87	91	**96**	77	76	85	**74**
#4	160	68	77	88	**97**	61	68	75	**89**
#5	110	48	79	92	**98**	37	61	82	**84**
#6	131	89	81	73	**98**	80	70	48	**88**
#7	153	80	90	88	**93**	60	76	65	**88**
All	1009	77	82	86	**97**	66	65	71	**83**

*NumV = Number of real Vehicles in the test set

that they provide complex and valid solutions for the vehicle detection task in general urban environments, we have also observed a number of limitations for each case. Most of the problems with *DEM-PCA* originate from the inaccuracies and discretization artifacts of the estimated elevation maps. In addition, short vegetation or various street objects can corrupt the process since their elevation range is often overlapping with the vehicles' height values. By testing the *h-max* method, we have noticed similar limitations as mentioned by the authors in [154]: in parking areas and cluttered regions, the technique yields inaccurate contours and merges some of the nearby objects, while vegetation causes a number of additional false alarms. Regarding the *Floodfill* algorithm, we observed that 3D connected component propagation is sensitive to noise due to partial occlusion, and nearby vehicles are often merged together. On the other hand, in the proposed technique the 2D projection implements already a noise filtering step, and the inverse object description approach of MPP does not request strictly connected components for detecting a vehicle. The quantitative comparison results shown in Table 5.3 confirm again the superiority of using MPP at the parent object level.

The *scooping detection* problem investigated in the third application example is a strongly technology specific issue, which understandably does not have a wide bibliography. However, the gain obtained by the *stochastic parent-child relationship* model of the EMPP can be well demonstrated in the context of this PCB inspection application. As a baseline technique for scooping detection, we have implemented a morphology-based solution called *Morph* (introduced in [11]), which applies two thresholding operations on the input image: The first one uses a lower threshold value yielding a binary solder paste candidate mask. Using the second threshold we extract the brightest image parts which are supposed to contain the scoop center areas. Finally a verification process removes the false scoop candidates. Table 5.4 shows the scooping detection performance of the

Table 5.4: PCB inspection task: Comparison of the child level performance on scooping detection between the *Morph* technique and the proposed EMPP model

PCB insp. method	TP	FP	FN	F-score
Morph technique [11]	514	228	150	73%
Proposed EMPP	629	65	35	93%

Table 5.5: Average computational time and parent object number for sample images of the different application fields

	Built-in	Aerial Traffic	PCB insp.
Avg. EMPP time	17.8 sec	11.1 sec	21.7 sec
Avg. *s*MPP time	13.9 sec	9.1 sec	20.1 sec
Avg. obj.num.	110	136	100

deterministic *Morph* and the stochastic EMPP approach: 20% gain can be reported for EMPP at the child level.

For keeping the *computational time* of the iterative Multi-level Multiple Birth-Death-Maintenance (M^MBDM) optimization algorithm low, we applied an exponential temperature cooling strategy, and took the advantage of the Bottom-Up Stochastic Entity Proposal (BUSEP) process similarly as an extension of the feature-based birth process from Chapter 4, by using various application-dependent image desciptors [5, 8, 10]. This way, the algorithm converged quickly to a sub-optimal solution, which proved to be efficient in the selected application domains. For quantitative analysis of the processing speed, we ran our algorithms on a standard desktop computer, and for each application we calculated the average computational time on one test image, both for the EMPP and *s*MPP models. Results listed in Table. 5.5 confirm that the EMPP's average running time varies between 11 and 22 seconds, which means a 20-30% computational overload versus *s*MPP for the built-in area analysis and aerial traffic surveillance tasks, while the running time of the two methods have been nearly identical for PCB analysis. The experiments also showed that the computational time is nearly independent of the number of objects, but it is related to the pixel based area of the parent objects, which was larger for the building detection and PCB inspection tasks.

Note that the iterative M^MBDM algorithm contains a number of stochastic operations: in each main step random moves mutate the population, such as probabilistic birth, death, parameter change or movement between groups etc. Nevertheless, we experienced that the outputs of the proposed framework are stable, i.e. the output configurations are largely similar for each run. In addition we have also performed a detailed analysis on the repeatability of the algorithm using an aerial Lidar segment containing 169 vehicles classified into 10 object groups. 200 independent

Table 5.6: Experiment repeatability for the vehicle detection task: Mean values and standard deviations of the measured error rates for 200 independent run in the same aerial Lidar segment

	TP	FP	FN	PFR	TG	FG
Mean	161.4	4.27	7.56	0.78	158.5	2.89
Dev	0.81	0.45	0.81	0.0077	2.37	2.24

Table 5.7: Distribution of the number of falsely grouped objects (out of 169 vehicles) in the 200-run experiment of Table 5.6

FG val.	0	1	2	3	4	5	6	7-20	21+
Freq.	26	36	20	41	41	25	8	3	0

experiments have been preformed on the same data and with the same parameter settings, and the output configurations of the stochastic method have been compared to the GT each time. Mean values and standard deviations of the measured error rates are shown in Table 5.6. We can observe that at the level of parent object recognition the deviations of TP/FN/FP are less than 1 object, while regarding the pixel-based rates it is less than 0.01 over the 200 test runs. As for object grouping, this scenario was one of the most challenging of all, since due to the low resolution of the aerial Lidar, the true object dimensions and orientations were often difficult to extract from the local point cloud data, thus the introduced object level grouping features strongly effected the output result. Table 5.7 shows the distribution of the numbers of falsely grouped objects (FG) during the 200 trials: typically 0-5 errors were measured among the 169 objects, and we experienced an FG larger than 6 only in three cases, while the error factor was never larger than 20.

5.8 Conclusion of the chapter

This chapter proposed a novel Embedded Marked Point Process (EMPP) model for joint extraction of objects, object groups, and specific object parts from high resolution digital images. The efficiency of the approach has been tested in three different application domains, and Ground Truth data has been prepared and published to enable quantitative evaluation. Based on the obtained results, we can confirm that the proposed EMPP model is able to handle real world tasks from significantly different application areas, providing a Bayesian framework for multi-level image content interpretation.

Chapter 6

4D environment perception

In this chapter various new solutions are proposed for the analysis of 4D (i.e. dynamic 3D) environment using up-do-date sensor configurations. The chapter begins with an overview on current challenges of environment perception and the opportunities provided by the latest sensor developments. Thereafter we define three different problems, which will be discussed in the remaining parts of the chapter. First, a new Marked Point Process approach is presented for 3D people localization and height estimation in multi-camera systems. Second, we introduce a new people surveillance framework based on point cloud sequences of a Rotating Multi-beam (RMB) Lidar sensor, with novel contributions in foreground-background separation of the RMB Lidar data streams, and person re-identification via Lidar based gait features. Third, we propose a workflow and several new algorithms for real time environment perception relying on a moving car-mounted RMB Lidar sensor, where as reference background map we use very dense 3D point clouds of the environment obtained by mobile laser scanning.

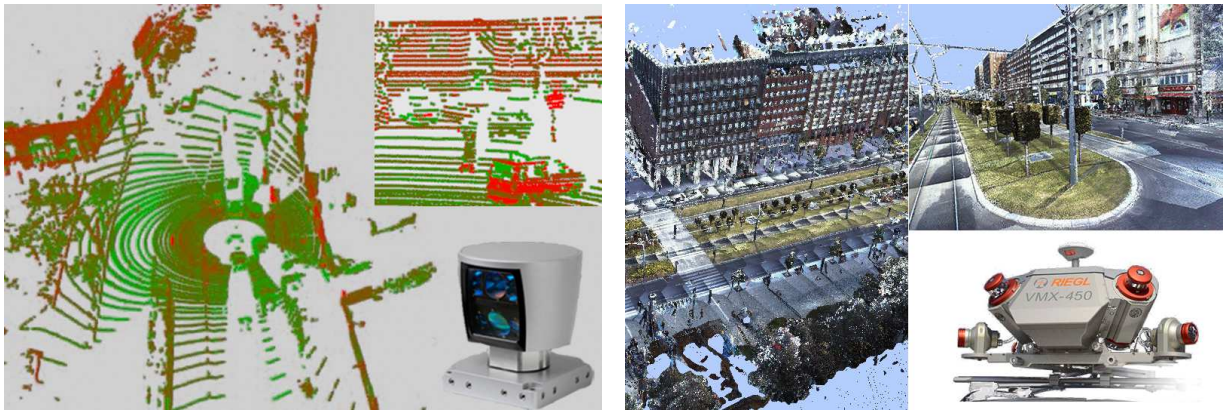
6.1 Introduction to 4D environment perception

Automated perception and interpretation of the surrounding environment are key issues in intelligent city management, traffic monitoring and control, security surveillance, or autonomous driving. Critical tasks involve detection, recognition, localization and tracking of various moving and static objects, environmental change detection and change classification. Nowadays a standard expectation is that the localization and tracking must be performed in the real 3D world coordinate system of the observed environment, which requirement – considering the temporal dimension of the measurements – implies 4D perception problems [27].

A significant part of the existing environment monitoring systems use electro-optical cameras as perception sensors, due to their established technologies, wide choices of the available properties and scalable prices. Nevertheless, despite the well explored literature of the topic, event analysis in optical image sequences may be still challenging in cases of crowded outdoor scenes due to uncontrolled illumination conditions, irrelevant background motion, and occlusions caused by various moving and static scene objects [155, 156]. In such situations multi-camera configurations can provide better solutions, since they monitor a dynamic scene from multiple viewpoints by taking the advantages of stereo-vision to exploit depth information for 3D localization and tracking [157, 158]. However, both mono and multi-camera systems suffer from a number of basic problems, such as artifacts due to moving shadows and low contrast between different objects in the color domain [159, 160], which issues raise still open research challenges in the topic.

As alternative solutions of conventional optical video cameras, range sensors offer significant advantages for scene analysis, since direct geometrical information is provided by them [50]. Using infra light based Time-of-Flight (ToF) cameras [161] or laser based Light Detection and Ranging (Lidar) sensors [162] enable recording directly measured range images, where we can avoid artifacts of the stereo vision based depth map calculation. From the point of view of data analysis, ToF cameras record depth image sequences over a regular 2D pixel lattice, where established image processing approaches, such as Markov Random Fields (MRFs) can be adopted for smooth and observation consistent segmentation and recognition [14]. However, such cameras can only be reliably used indoors, due to limitations of current infra-based sensing technologies, and usually they have a limited Field of View (FoV), which fact can be a drawback for surveillance and monitoring applications.

Rotating Multi-beam (RMB) Lidar systems provide a 360° FoV of the scene, with a vertical resolution equal to the number of the sensors, while the horizontal angle resolution depends on the speed of rotation (see Fig. 6.1(a)). Each laser point of the output point cloud is associated with 3D spatial coordinates, and possibly with auxiliary channels such as reflection number or an intensity value of laser reflection. RMB Lidars can produce high frame-rate *point cloud videos* enabling



(a) Velodyne HDL-64E RMB Lidar (b) Riegl VMX-450 MLS

Figure 6.1: Data comparison of two different Lidar sensors: (a) a time frame from a dynamic RBM Lidar sequence and (b) static point cloud scene obtained by Mobile Laser Scanning (MLS)

dynamic event analysis in the 3D space. On the other hand, the measurements have a low spatial density, which quickly decreases as a function of the distance from the sensor, and the point clouds may exhibit particular patterns typical to sensor characteristic (such as ring patterns in Fig. 6.1(a)). Although the 3D measurements are quite accurate (up to few *cm*s) in the sensor's local coordinate system, the global positioning error of the vehicles may reach several meters due to limitations of the availability of external navigation signals.

Mobile laser scanning (MLS) platforms equipped with time synchronized Lidar sensors and navigation units can rapidly provide very dense and accurate point clouds from large environments (see Fig. 6.1(b)), where the 3D spatial measurements are accurately registered to a geo-referenced global coordinate system [163, 164, 165]. In the near future, these point clouds may act as a basis for detailed and up-to-date 3D High Definition (HD) maps of the cities, which can be be utilized by self driving vehicles for navigation, or by city authorities for road network management and surveillance, architecture or urban planning. While the high speed of point cloud acquisition is a clear advantage of MLS, due to the huge data size yielded by each daily mission, applying efficient automated data filtering and analyzing algorithms in the processing side is crucially needed.

In this chapter, we introduce various contributions in the 4D environment perception topic. Sec. 6.2 presents a new method on Marked Point Process (MPP) based pedestrian localization and height estimation in multi-camera systems, and gives a detailed comparative evaluation of the proposed method versus a state-of-the art technique. In Sec. 6.3 we introduce a new 4D people surveillance framework, with original methodological contributions in motion detection, gait-based pedestrian re-identification and activity recognition using a single RMB Lidar sensor which monitors the scene from a fixed position. Finally, in Sec. 6.4 we propose a new workflow

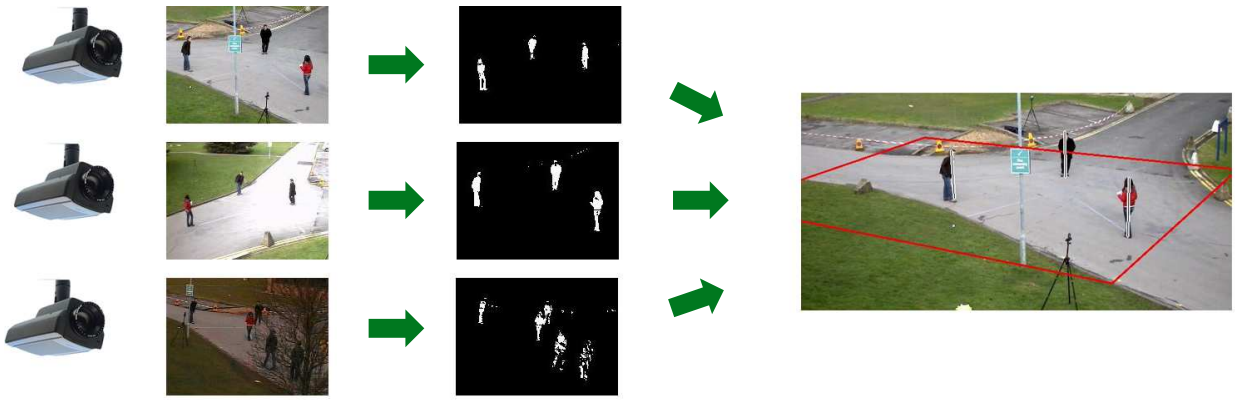


Figure 6.2: Multiview people detection and height estimation

and various specific algorithms for dynamic urban scene analysis with a car-mounted moving RMB Lidar sensor, exploiting MLS data as a HD background map.

6.2 People localization in multi-camera systems

Person localization is a crucial step in people surveillance applications, since it is an important precursor of tracking and activity analysis. At each time frame, the 3D ground positions of the observed pedestrians should be automatically extracted in the world coordinate system. A possible approach for the problem is using multi-camera systems, which are able to monitor the scene from multiple viewpoints simultaneously, providing the advantage that people partially occluded from certain viewpoints might be clearly observable from another ones. Here people detection and localization require 3D information retrieval from stereo or multi-view inputs, with efficient approximation strategies of the missing information resulted by camera noise, artifacts of image matching (especially in featureless regions) and occlusion [36].

In this section we introduce a Bayesian approach on multiple people localization in multi-camera systems [39]. First, pixel-level features are extracted, which are based on physical properties of the 2D image formation process, and provide information about the head and leg positions of the pedestrians, distinguishing standing and walking people, respectively [41]. Then features from the multiple camera views are fused to create evidence for the location and height of people in the ground plane. This evidence accurately estimates the leg position even if either the area of interest is only a part of the scene, or the overlap ratio of the silhouettes with irrelevant background motion within the monitored area is significant. Using this sort of information we create a 3D object configuration model. We also utilize prior geometrical constraints, which describe the possible interactions between two pedestrians. To approximate the position of the people, we use a

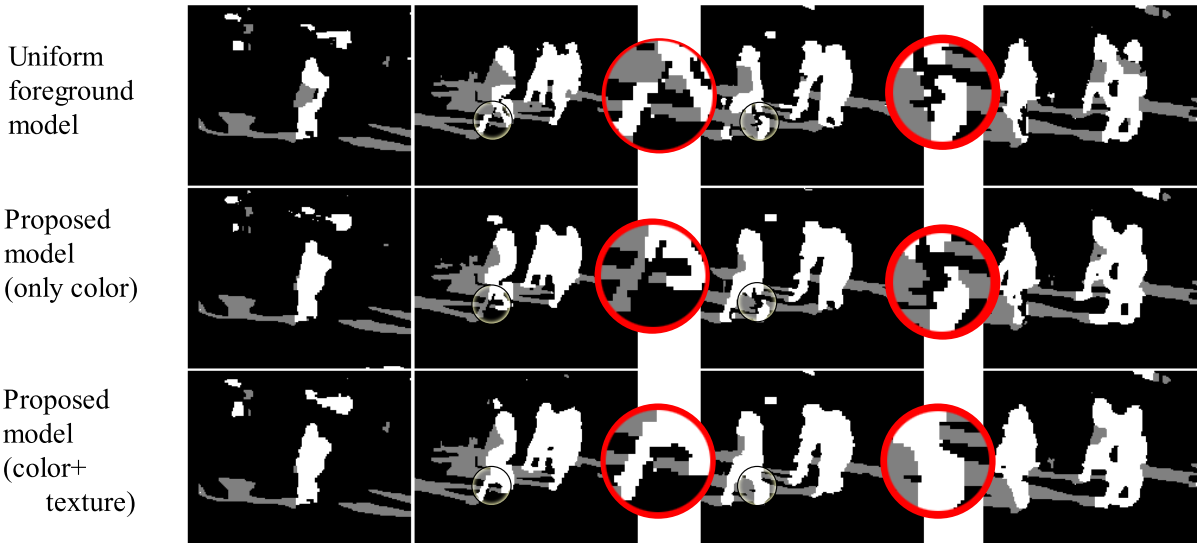


Figure 6.3: *Foreground detection results* with our approach [14] and two reference techniques on a real surveillance video sequence. White, gray and black pixel mark foreground, shadow and background classes. Row 1: MRF with the reference 'uniform foreground' calculus [166]. Row 2: Our initially proposed model [47] without using microstructural features. Row 3: Segmentation results with our final model [14].

population of 3D cylinder objects, which is realized by a Marked Point Process. The final configuration results are obtained by an iterative stochastic energy optimization algorithm. The proposed approach is evaluated on two publicly available datasets, and compared to a recent state-of-the-art technique. To obtain relevant quantitative test results, a 3D Ground Truth annotation of the real pedestrian locations is prepared, while two different error metrics and various parameter settings are proposed and evaluated, showing the advantages of our proposed model.

6.2.1 A new approach on multi-view people localization

The first step of the workflow is foreground detection in the video frames of each camera. For this purpose, we use our earlier proposed Markov Random Field (MRF) based approach [14], which considers the task as a three-class segmentation problem with *foreground*, *background* and *moving shadow* classes[1]. The applied technique has four key features [14]: (i) It uses a new *parametric* shadow model [51], where local feature vectors are derived at the individual pixels, and the shadow's domain is represented by a global probability density function in that feature space. The parameter adaption algorithm is based on following the changes in the shadow's feature domain.

[1]The proposed foreground detection approach has already been presented in the Ph.D. dissertation of the Author [53], therefore, we only give here a short overview. The Reader may find the details in the corresponding publications [14, 15, 51, 53].

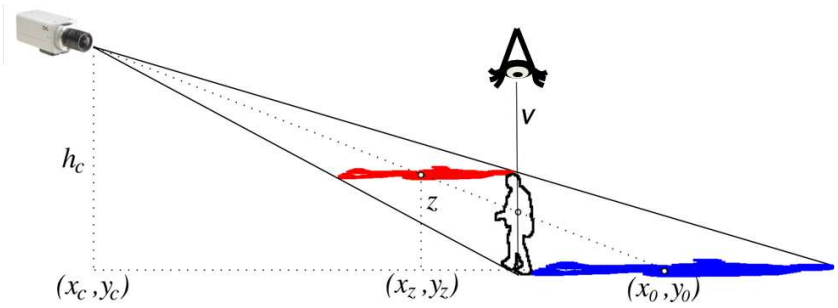
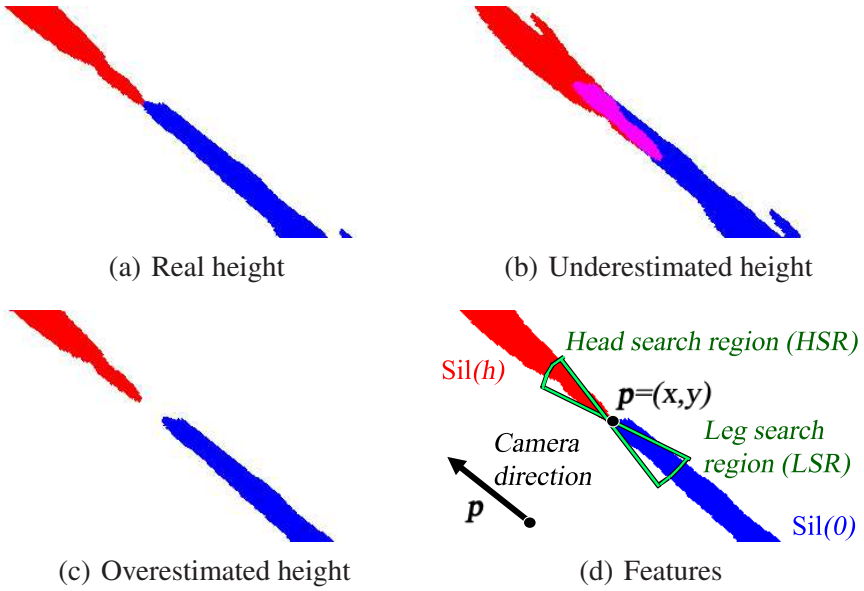


Figure 6.4: Side view sketch of a person's silhouette projected to the ground plane (blue) and to the horizontal plane intersecting the top of the head (red).

(ii) Our model encapsulates a novel multi-modal color model for the foreground class, which exploits spatial color statistics instead of high frame rate temporal information to describe the regions of moving objects. Using the assumption that any object consists of spatially connected parts which have typical color/texture patterns, the distribution of the likely foreground colors have been locally estimated in each pixel neighborhood. (iii) We have developed a probabilistic description of microstructural responses observed in the background and in shadows, where the features can be defined by arbitrary 3×3 kernels. At different pixel positions different kernels could be used, and an adaptive kernel selection strategy has been proposed considering the local textural properties of the background regions. (iv) We have also prepared a detailed experimental comparison of various widely used color spaces in connection with our proposed framework [15]. We have shown that color space selection is a key issue in shadow detection, if for practical purposes, shadow models with less free parameters are preferred, and the experiments confirmed the clear superiority of the CIE L*u*v* color space. Some example results with the proposed model, and competing approaches are shown in Fig. 6.3.

The input of the remaining steps in the proposed multi-view person localization method consists of the foreground masks extracted from multiple calibrated camera views [167], monitoring the same scene. The main idea of our method is to project the extracted foreground pixels both on the ground plane, and on the horizontal plane shifted to the height of the person (see Fig. 6.4). This projection will create a distinct visual feature, observable from a virtual birds-eye viewpoint above the ground plane. However, the person's height is unknown *a priori*, and the height of different people in the scene may also be different. Therefore, we project the silhouette masks on multiple parallel planes at heights in the range of typical human height. In crowded scenes the overlap rate between person silhouettes in the individual foreground masks is usually high, which could corrupt our hypothesis. We solve this problem by fusing the projected results of multiple camera views on the same planes. Finally, we search for the optimal configuration through *stochastic optimization* using the extracted features and geometrical constraints.



(a) Real height (b) Underestimated height

(c) Overestimated height (d) Features

Figure 6.5: Feature definition

6.2.2 Silhouette based feature extraction

Let us denote by P_0 the groundplane, and by P_z the parallel plane above P_0 with an elevation z. In the first step of the proposed method we project the detected silhouettes to P_0 and to different P_z planes (with different $z > 0$ offsets) by using the projection model of the calibrated cameras. Consider the person with height h presented in Fig 6.4, where we projected the silhouette on the P_0 ground plane (marked with blue) and the P_z plane with the height of the person (ie. $z = h$, marked with red). Also consider the v vertical axis of the person which is perpendicular to the P_0 plane. We can observe that from this axis, the silhouette points projected to the $P_z|_{z=h}$ plane lie in the direction to the camera, while the silhouette print on P_0 is on the opposite side of v. For more precise investigations, in Fig. 6.5 the scene is visualized from a viewpoint above P_z, watching down in a perpendicular direction to the ground. Here, the silhouette prints from P_z and P_0 are projected to a common $x - y$ plane and jointly shown by red and blue colors, respectively (overlapping areas are purple). We can observe in Fig. 6.5(a), that if the height estimation is correct ($z = h$), the two prints just touch each other in the $p = (x, y)$ point which corresponds to the ground position of the person. However, if the height of P_z is underestimated (i.e. $z < h$), the two silhouette prints will overlap as shown in Fig. 6.5(b). When the height is overestimated (ie. $z > h$), the silhouettes will move away, see Fig. 6.5(c).

Next, we derive a fitness function which evaluates the hypothesis of a proposed scene object with ground position $\mathbf{p} = (x, y)$ and height h, using the multiple camera information. As shown in Fig. 6.5(d), for a given camera projection we can define for each $\mathbf{p}$ position candidate a head

search region (HSR($\mathbf{p}$)) and a leg search region (LSR($\mathbf{p}$)), denoted by green circle sectors. If both the $\mathbf{p}$ ground position and the h height estimations are accurate, we expect several P_h-based (red) silhouette points (Sil(h)) in HSR but not in LSR. Regarding the P_0 (blue) silhouette points (Sil(0)) our expectation is the opposite: low coverage in HSR and high coverage in LSR. This observation leads to the following fitness features at the ith camera view:

$$f_{hd}^i(\mathbf{p},h) = \frac{\mathbf{Area}\big(\mathrm{Sil}^i(h) \cap \mathrm{HSR}^i(\mathbf{p})\big) - \mathbf{Area}\big(\mathrm{Sil}^i(h) \cap \mathrm{LSR}^i(\mathbf{p})\big)\big)}{\mathbf{Area}\big(\mathrm{HSR}^i(\mathbf{p})\big)} \,. \tag{6.1}$$

$$f_{lg}^i(\mathbf{p}) = \frac{\mathbf{Area}\big(\mathrm{Sil}^i(0) \cap \mathrm{LSR}^i(\mathbf{p})\big) - \mathbf{Area}\big(\mathrm{Sil}^i(0) \cap \mathrm{HSR}^i(\mathbf{p})\big)\big)}{\mathbf{Area}\big(\mathrm{LSR}^i(\mathbf{p})\big)} \,. \tag{6.2}$$

If the object defined by the $[\mathbf{p},h]$ parameters is completely visible for the ith camera, both the $f_{hd}^i(\mathbf{p},h)$ and $f_l^i(\mathbf{p})$ features should have *high* values. However, in the available views, some of the legs or heads may be partially or completely occluded by other pedestrians or static scene objects, which can strongly corrupt the feature values. Although the descriptors may be weak in the individual cameras, we can construct a stronger feature if we average the responses of the N available cameras, *i.e.*

$$\bar{f}_{hd}(\mathbf{p},h) = \frac{1}{N} \cdot \sum_{i=1}^{N} f_{hd}^i(\mathbf{p},h) \,, \quad \bar{f}_{lg}(\mathbf{p}) = \frac{1}{N} \cdot \sum_{i=1}^{N} f_{lg}^i(\mathbf{p}) \,. \tag{6.3}$$

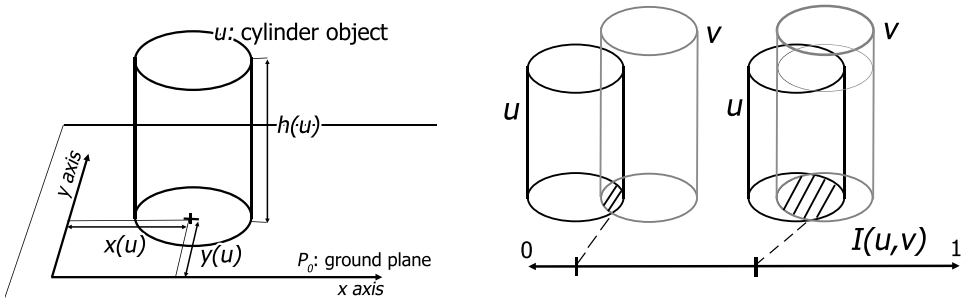
Finally, the joint data feature $f(\mathbf{p},h)$ is derived as

$$f(\mathbf{p},h) = \sqrt{\bar{f}_{hd}(\mathbf{p},h) \cdot \bar{f}_{lg}(\mathbf{p})} \,, \tag{6.4}$$

6.2.3 3D Marked Point Process model

The $f(\mathbf{p},h)$ feature introduced in the previous section evaluates the hypothesis that a person with a height h stands in the ground position $\mathbf{p}$, based on the multiple camera measurements. Our goal is to recognize a configuration of an unknown number of people in the scene, where each person is characterized by the (x,y,h) parameter triplet with $\mathbf{p} = (x,y)$. Since this problem can be formulated as a population extraction task, we have embedded the features into a Marked Point Process (MPP) model. While in MPP solutions in Chapters 4 and 5 of this Thesis the objects were represented by various 2D geometric figures, in this case, a 3D cylinder model describes a given person (see Fig. 6.6(a)).

We monitor a rectangular Region of Interest (RoI) on P_0 discretized into $S_W \times S_H$ locations corresponding to a regular grid, and also round the person heights to integers measured in cm. Therefore, the object space $\mathcal{H}$ can be obtained as $\mathcal{H} = [1,\dots,S_W] \times [1,\dots,S_H] \times [h_{\min},\dots,h_{\max}]$.



(a) Cylinder model (b) Intersection feature

Figure 6.6: Cylinder objects modeling people in the 3D scene coordinate system. Their ground plane position and height will be estimated. Intersection of cylinders in the 3D space is used as geometrical constraint in the object model

The remaining part of the MPP model construction follows the description in Sec. 2.2.1. An object u is described by the (x, y, h) parameters, since we used cylinders with a fixed R radii corresponding to the minimal expected half-distance between two ground positions. The $A(u)$ data terms are derived from the $f(\mathbf{p}, h)|_{\mathbf{p}=(x,y)}$ feature values using the nonlinear $\mathcal{M}(f, d_0, D)$ feature mapping function of formula (4.1). The $I(u, v)$ interaction terms prescribe a non-overlapping constraint between neighboring cylinders, as demonstrated in Fig. 6.6(b):

$$I(u, v) = \frac{\mathbf{Volume}\big(u \cap v\big)}{\mathbf{Volume}\big(u \cup v\big)} \,. \tag{6.5}$$

Finally, the Multiple Birth and Death optimization technique (introduced in Sec. 2.2.3) is utilized to obtain the targeted population.

6.2.4 Evaluation of multi-camera people localization

We have compared our approach to the Probabilistic Occupancy Map (POM) technique [157], which has been a state-of-the-art method with similar purposes[1].

For the evaluation of the two methods we used two public sequences. First, from the PETS 2009 dataset [168] we selected the *City center* images containing approximately 1 minute of recordings (400 frames total) in an outdoor environment. From the available views we selected cameras with large fields of view (View_001, View_002, and View_003) and we used a RoI of size $12.2\text{m} \times 14.9\text{m}$, which is visible from all three cameras. The maximum number of pedestrians at the same time inside the RoI is 8.

[1]Executable application of the POM reference technique is freely available (on 26.04.2019) at `http://cvlab.epfl.ch/software/pom/`

Figure 6.7: Detection examples by the proposed 3DMPP model in the *City Center* sequence with multiple pedestrians and occlusions, projected to one of the camera images (note: as discussed, the detection is based on multiple camera views)

The second dataset we used in our experiments is the EPFL *Terrace* dataset, which is 3 minutes and 20 seconds long (5000 frames total). The scene is semi-outdoor, since it was recorded in a controlled outdoor environment and it also lacks some important properties of a typical outdoor scene (*e.g.* no background motion caused by the moving vegetation is present, and no static background objects occlude the scene). We selected three cameras having small fields of view, and defined the RoI as a $5.3\text{m} \times 5.0\text{m}$ rectangle. The scene is severely cluttered in some periods.

For numerical evaluation we created complete ground position annotation for both the *City center* and *Terrace* multi-camera sequences, using a newly developed 3D Ground Truth annotations tool introduced in [35]. For the *City center* sequence we annotated all 400 frames, while the *Terrace* sequence has been annotated with 1Hz frequency resulting in 200 annotated frames

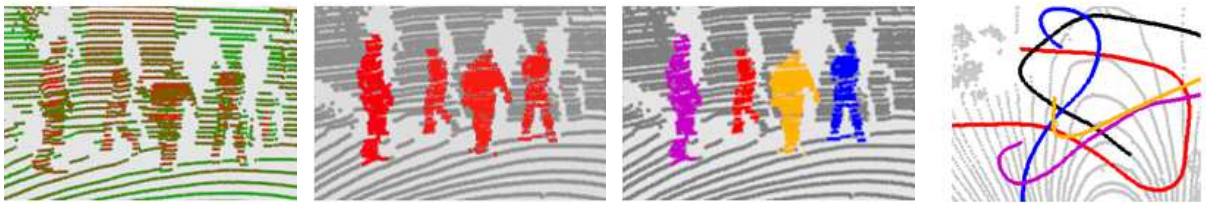
We considered various error rates: *Missed Detections (MD)* counts the number of examples, where no detection could have been assigned to a Ground Truth target. *False Detections (FD)* value corresponds to examples, where no Ground Truth position could have been assigned to a detected sample. *Multiple Instances (MI)* measures the number of cases where multiple detections were assigned to a single Ground Truth position. Finally, the *Total Error (TE)* is taken as $\mathrm{TE} = \mathrm{MD} + \mathrm{FD} + \mathrm{MI}$.

After counting all the false localization results (MD, FD, MI) on all annotated frames we express them in percent of the number of all objects, and we denote these ratios by MDR, FDR, MIR, and TER. Note that while $\mathrm{MDR} \leq 1$ and $\mathrm{MIR} \leq 1$ always hold, in case of many false alarms FDR (thus also TER) may exceed 1. For accuracy evaluation of position estimation, we also measured distance between the ground positions in the Ground Truth annotation and in the detection results yielding the Ground Position Error (GPE) metric.

Detection examples by the proposed model in the two sample frames of the *City center* sequence are displayed in Fig. 6.7. We can numerically compare POM to the proposed 3DMPP method in Table 6.1, considering both test sequences and the GPE error metrics. Here in all cases

Table 6.1: Comparison of the POM and the proposed 3DMPP models with optimized parameter sets (so that the total error rate TER is minimized), all three cameras are used

Sequence	Method	Ground Position Errors (GPE)			
		TER	FDR	MDR	MIR
City center	POM	**0.252**	0.179	0.073	0.000
	Prop. 3DMPP	**0.122**	0.020	0.096	0.006
Terrace	POM	**0.686**	0.354	0.331	0.001
	Prop. 3DMPP	**0.131**	0.043	0.083	0.005



(a) Lidar sensor (b) Foreground detection (c) Person separation (d) Top view trajectories

Figure 6.8: Lidar based surveillance: flowchart of the motion detection and person tracking process

the parameters have been set to minimize TER, while the corresponding FDR, MDR and MIR values are also listed. Results confirm the superiority of the proposed 3DMPP model over POM. A detailed study on parameter sensitivity of the proposed model has also been provided in [9].

6.3 A Lidar based 4D people surveillance approach

This section presents new approaches for people surveillance based on data streams of a Rotating Multi-beam (RMB) Lidar sensor standing in a static position. First we deal with efficient motion detection and tracking (see Fig. 6.8), thereafter we focus on gait based person re-identification during the surveillance period, as well as recognition of specific activity patterns. *Main contributions* of this sections are twofold. On one hand we introduce a new dynamic MRF model for foreground-background segmentation of RMB Lidar streams, and show its superiority versus straightforward approaches. On the other hand we propose machine learning based techniques for gait and activity analysis on the Lidar data, assuming that the descriptors for training and recognition are observed and extracted from realistic outdoor surveillance scenarios, where multiple pedestrians are walking in the field of interest following possibly intersecting trajectories, thus the observations might often be affected by occlusions or background noise.

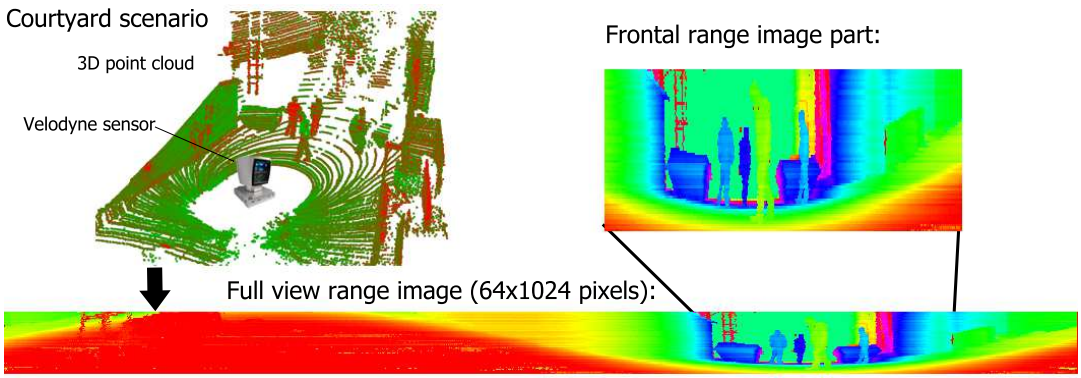


Figure 6.9: Point cloud recording and range image formation with a Velodyne HDL-64E RMB Lidar sensor

6.3.1 Foreground extraction in Lidar point cloud sequences

A Rotating Multi-beam (RMB) Lidar sensor provides a time sequence of 3D point clouds capturing a 360° FoV of the scene. For efficient data processing, the 3D RMB Lidar points are often projected onto a cylinder shaped range image [162, 169] as shown in Fig. 6.9. However, this mapping is usually ambiguous: On one hand, several laser beams with slight orientation differences are assigned to the same pixel, although they may return from different surfaces. As a consequence, a given pixel of the range image may represent different background objects at the consecutive time steps.

In this section we introduce a hybrid 2D–3D approach [6, 31] for dense foreground-background segmentation of RMB Lidar point cloud sequences obtained from a fixed sensor position (Fig. 6.9 and 6.10). Our technique solves the computationally critical spatial filtering steps in the 2D range image domain by an MRF model, however, ambiguities of discretization are handled by joint consideration of true 3D positions and back projection of 2D labels. By developing a spatial foreground model, we significantly decrease the spurious effects of irrelevant background motion, which principally caused by moving tree crowns and bushes.

6.3.1.1 Problem formulation and data mapping

Assume that the RMB Lidar system contains R vertically aligned sensors, and rotates around a fixed axis with a possibly varying speed[1]. The output of the Lidar within a time frame t is a *point cloud* of $l^t = R \cdot c^t$ points: $\mathcal{L}^t = \{p_1^t, \ldots, p_{l^t}^t\}$. Here c^t is the number of point *columns* obtained at t, where a given column contains R concurrent measurements of the R sensors, thus c^t depends on the rotation speed. Each point, $p \in \mathcal{L}^t$, is associated to sensor distance $d(p) \in [0, D_{\max}]$,

[1]The speed of rotation can often be controlled by software, but even in case of constant control signal, we must expect minor fluctuations in the measured angle-velocity, which may result in different number of points for different 360° scans in time.

pitch index $\hat{\vartheta}(p) \in \{1, \ldots, R\}$ and yaw angle $\varphi(p) \in [0, 360°]$ parameters. $d(p)$ and $\hat{\vartheta}(p)$ are directly obtained from the Lidar's data flow, by taking the measured distance and sensor index values corresponding to p. Yaw angle $\varphi(p)$ is calculated from the Euclidean coordinates of p projected to the ground plane, since the R sensors have different horizontal view angles, and the angle correction of calibration may also be significant [170].

For efficient data manipulation, we also introduce a range image mapping of the obtained 3D data. We project the point cloud to a cylinder, whose central basis point is the ground position of the RMB Lidar and the axis is prependicular to the ground plane. Note that slightly differently from [169], this mapping is also efficiently suited to configurations, where the Lidar axis is tilted to increase the vertical Field of View. Then we stretch a $S_H \times S_W$ sized 2D pixel lattice S on the cylinder surface, whose height S_H is equal to the R sensor number, and the width S_W determines the fineness of discretization of the yaw angle. Let us denote by s a given pixel of S, with $[y_s, x_s]$ coordinates. Finally, we define the $\mathcal{P} : \mathcal{L}^t \to S$ point mapping operator, so that y_s is equal to the pitch index of the point and x_s is set by dividing the $[0, 360°]$ domain of the yaw angle into S_W bins:

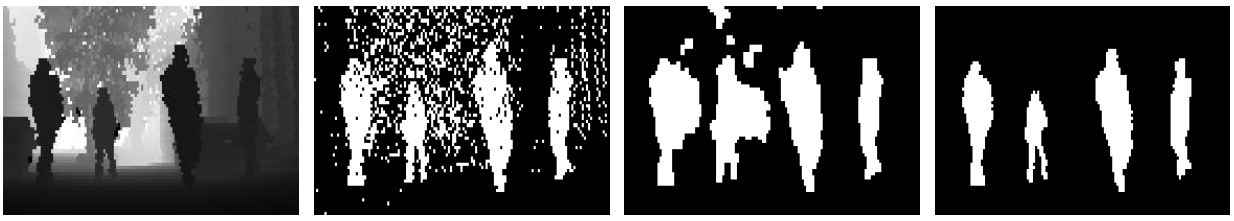
$$s \stackrel{\text{def}}{=} \mathcal{P}(p) \text{ iff } y_s = \hat{\vartheta}(p),\ x_s = \text{round}\left(\varphi(p) \cdot \frac{S_W}{360°}\right) \tag{6.6}$$

The goal of the foreground detector module is at a given time frame t to assign each point $p \in \mathcal{L}^t$ to a label $\varsigma(p) \in \{\text{fg}, \text{bg}\}$ corresponding to the moving object (i.e. foreground, fg) or background classes (bg), respectively.

6.3.1.2 Background model

The background modeling step assigns a fitness term $f_{\text{bg}}(p)$ to each $p \in \mathcal{L}^t$ point of the cloud, which evaluates the hypothesis that p belongs to the background. The process starts with a cylinder mapping of the points based on eq. (6.6), where we use a $R \times S_W^{\text{bg}}$ pixel lattice S^{bg} (R is the sensor number). For each s cell of S^{bg}, we maintain a Mixture of Gaussians (MoG) approximation of the $d(p)$ distance histogram of p points being projected to s. Following the approach of [171], we use a fixed K number of components (here $K = 5$) with weight κ_s^i, mean μ_s^i and standard deviation σ_s^i parameters, $i = 1 \ldots K$. Then we sort the weights in decreasing order, and determine the minimal k_s integer which satisfies $\sum_{i=1}^{k_s} \kappa_s^i > T_{\text{bg}}$ (we used here $T_{\text{bg}} = 0.89$). We consider the components with the k_s largest weights as the background components. Thereafter, denoting by $\eta()$ a Gaussian density function, and by $\mathcal{P}^{\text{bg}}$ the projection transform onto S^{bg}, the $f_{\text{bg}}(p)$ background evidence term is obtained as:

$$f_{\text{bg}}(p) = \sum_{i=1}^{k_s} \kappa_s^i \cdot \eta\left(d(p), \mu_s^i, \sigma_s^i\right),\ \text{where } s = \mathcal{P}^{\text{bg}}(p). \tag{6.7}$$



(a) Range image part (b) Basic MoG [162] (c) uniMRF [160] (d) Proposed DMRF

Figure 6.10: Foreground segmentation in a range image part with three different methods

The Gaussian mixture parameters are set and updated based on [171], while we used $S_W^{\mathrm{bg}} = 2000$ angle resolution, which provided the most efficient detection rates in our experiments. By thresholding $f_{\mathrm{bg}}(p)$, we can get a dense foreground/background labeling of the point cloud [162, 171] (referred later as *Basic MoG* method), but as shown in Fig. 6.10(b), this classification is notably noisy in scenarios recorded in large outdoor scenes.

6.3.1.3 DMRF approach on foreground segmentation

In this section, we propose a Dynamic Markov Random Field (DMRF) model to obtain smooth and observation consistent segmentation of the point cloud sequence (Fig. 6.11). Since MRF optimization in 3D is computationally expensive, we define the DMRF model in the range image space, and 2D image segmentation is followed by a point classification step to handle ambiguities of the mapping. As defined by eq. (6.6), we use a $\mathcal{P}$ cylinder projection transform to obtain the range image, with a $S_W = \hat{c} < S_W^{\mathrm{bg}}$ grid width, where $\hat{c}$ denotes the expected number of point columns of the point sequence in a time frame. By assuming that the rotation speed is slightly fluctuating, this selected resolution provides a dense range image, where the average number of points projected to a given pixel is around 1. Let us denote by $P_s \subset \mathcal{L}^t$ the set of points projected to pixel s. For a given direction, foreground points are expected being closer to the sensor than the estimated mean background range value. Thus, for each pixel s we select the closest projected point $p_s^t = \mathrm{argmin}_{p \in P_s} d(p)$, and assign to pixel s of the range image the $d_s^t = d(p_s^t)$ distance value. For 'undefined' pixels ($P_s = \emptyset$), we interpolate the distance from the neighborhood. For spatial filtering, we use an eight-neighborhood system in S, and denote by $\mathcal{N}_s \subset S$ the neighbors of s.

Next, we assign to each $s \in S$ foreground and background energy (i.e. negative fitness) terms, which describe the class memberships based on the observed $d(s)$ values. The background energies are directly derived from the parametric MoG probabilities using (6.7): $\epsilon_{\mathrm{bg}}^t(s) = -\log\left(f_{\mathrm{bg}}(p_s^t)\right)$.

For description of the foreground, using a constant ϵ_{fg} could be a straightforward choice [160] (we call this approach *uniMRF*), but this uniform model results in several false alarms due to background motion and quantization artifacts. Instead of temporal statistics, we use spatial distance

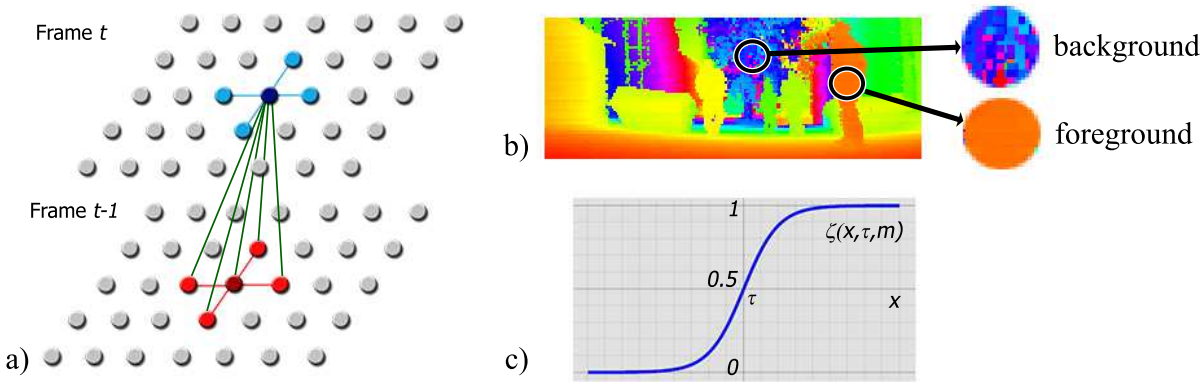


Figure 6.11: Components of the dynamic MRF model. a) structure of the multi-layer MRF model b) demonstrating the different local range value distributions in the neighborhood of a given foreground and background pixel, respectively c) plot of the used sigmoid function.

similarity information to overcome this problem by using the following assumption: whenever s is a foreground pixel, we should find foreground pixels with similar range values in the neighborhood (Fig. 6.11(b)). For this reason, we use a non-parametric kernel density model for foreground:

$$\epsilon^t_{\mathrm{fg}}(s) = \sum_{r \in \mathcal{N}_s} \zeta(\epsilon^t_{\mathrm{bg}}(r), \tau_{\mathrm{fg}}, m_\star) \cdot k\left(\frac{d^t_s - d^t_r}{h}\right),$$

where h is the kernel bandwidth and $\zeta : \mathbb{R} \to [0,1]$ is a sigmoid function (see Fig. 6.11(c)):

$$\zeta(x, \tau, m) = \frac{1}{1 + \exp(-m \cdot (x - \tau))}. \tag{6.8}$$

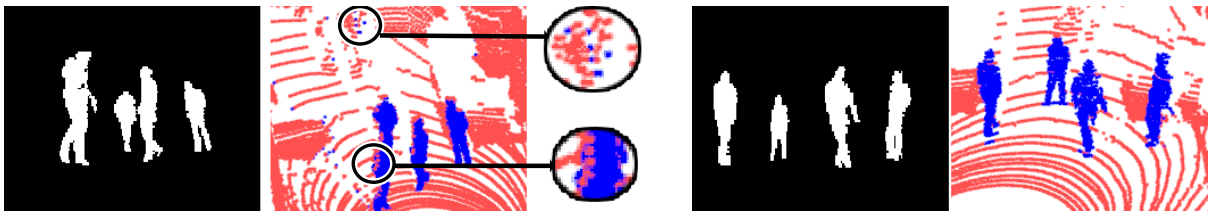
We use here a uniform kernel: $k(x) = \mathbf{1}\{|x| \leq 1\}$, where $\mathbf{1}\{.\} \in \{0,1\}$ is the binary indicator function of a given event.

To formally define the range image segmentation task, to each pixel $s \in S$, we assign a $\varsigma^t_s \in \{\mathrm{fg}, \mathrm{bg}\}$ class label so that we aim to minimize the following energy function:

$$E = \sum_{s \in S} V_D(d^t_s | \varsigma^t_s) + \sum_{s \in S} \underbrace{\sum_{r \in \mathcal{N}_s} \alpha \cdot \mathbf{1}\{\varsigma^t_s \neq \varsigma^{t-1}_r\}}_{\xi^t_s} + \sum_{s \in S} \underbrace{\sum_{r \in \mathcal{N}_s} \beta \cdot \mathbf{1}\{\varsigma^t_s \neq \varsigma^t_r\}}_{\chi^t_s}, \tag{6.9}$$

where $V_D(d^t_s|\varsigma^t_s)$ denotes the data term, while ξ^t_s and χ^t_s are the temporal and spatial smoothness terms, respectively, with $\alpha > 0$ and $\beta > 0$ constants. Let us observe, that although the model is dynamic due to dependencies between different time frames (see the ξ^t_s term), to enable real time operation, we develop a causal system, i.e. labels from the past are not updated based on labels from the future. The data terms are derived from the data energies by sigmoid mapping:

$$V_D(d^t_s | \varsigma^t_s = \mathrm{bg}) = \zeta(\epsilon^t_{\mathrm{bg}}(s), \tau_{\mathrm{bg}}, m_{\mathrm{bg}})$$
$$V_D(d^t_s | \varsigma^t_s = \mathrm{fg}) = \begin{cases} 1, & \text{if } d^t_s > \max_{\{i=1\ldots k_s\}} \mu^{i,t}_s + d_0 \\ \zeta(\epsilon^t_{\mathrm{fg}}(s), \tau_{\mathrm{fg}}, m_{\mathrm{fg}}), & \text{otherwise.} \end{cases}$$



(a) Naive backprojection (b) Proposed backprojection

Figure 6.12: Backprojection of the range image labels to the point cloud. (a) simple backprojection with assigning the same label to s and p, whenever $s = \mathcal{P}(p)$. (b) result of the proposed backprojection scheme

The sigmoid parameters τ_{fg}, τ_{bg}, m_{fg}, m_{bg} and $m_\star$ can be estimated by Maximum Likelihood strategies based on a few manually annotated training images. As for the smoothing factors, we use $\alpha = 0.2$ and $\beta = 1.0$ (i.e. the spatial constraint is much stronger), while the kernel bandwidth is set to $h = 30$cm. The MRF energy eq. (6.9) is minimized via the fast graph-cut based optimization algorithm [105].

The result of the DMRF optimization is a binary foreground mask on the discrete S lattice. As shown in Fig. 6.12, the final step of the method is the classification of the points of the original $\mathcal{L}$ cloud, considering that the projection may be ambiguous, i.e. multiple points with different true class labels can be projected to the same pixel of the segmented range image. With denoting by $s = \mathcal{P}(p)$ for time frame t, we use the following strategy:

- $\varsigma(p) = \text{fg}$, iff one of the following two conditions holds:
  $(\dagger)$ $\varsigma_s^t = \text{fg}$ and $d(p) < d_s^t + 2 \cdot h$
  $(\ddagger)$ $\varsigma_s^t = \text{bg}$ and $\exists r \in \mathcal{N}_r : \{\varsigma_r^t = \text{fg}, |d_r^t - d(p)| < h\}$
- $\varsigma(p) = \text{bg}$: otherwise.

The above constraints eliminate several $(\dagger)$ false positive and $(\ddagger)$ false negative foreground points, projected to pixels of the range image near the object edges, which improvement can be seen by comparing the left and right examples of Fig. 6.12.

6.3.2 Pedestrian detection and tracking

The next step is pedestrian detection and tracking. The input of this component is the RMB Lidar point cloud sequence, where each point is marked with a segmentation label of foreground or background, while the output consists of clusters of foreground regions so that the points corresponding to the same object receive the same label over the sequence.

First, the point cloud regions classified as foreground are clustered to obtain separate blobs for each moving person candidate. A regular lattice is fit to the ground plane and the foreground regions are projected onto this lattice. Morphological filters are applied in the image plane to

obtain spatially connected blobs for different persons. Then the system extracts appropriately sized connected components that satisfy area constraints determined by lower and higher thresholds. The centre of each extracted blob is considered as a candidate for foot position on the ground.

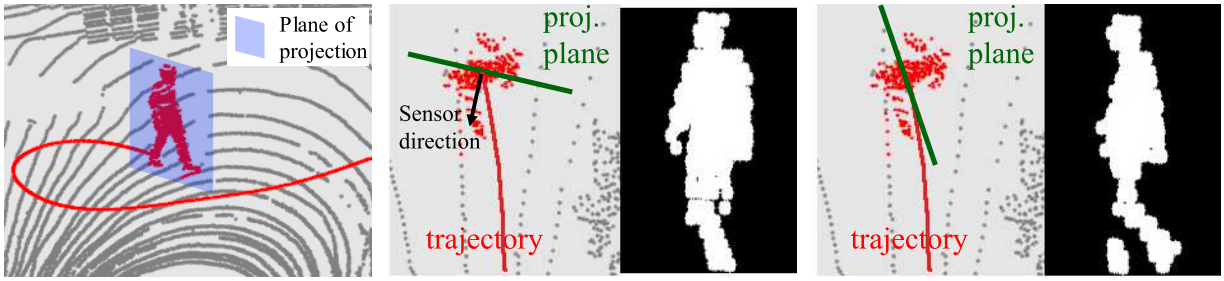
The pedestrian tracking module combines Short-Term Assignment (STA) and Long-Term Assignment (LTA) steps. The STA part attempts to match each actually detected object candidate with the current object trajectories maintained by the tracker, by purely considering the projected 2D centroid positions of the target. The STA process should also be able to continue a given trajectory if the detector misses the concerning object for a few frames due to occlusion. In these cases the temporal discontinuities of the tracks must be filled with estimated position values, as detailed in [6]. On the other hand, the LTA module is responsible for extracting discriminative features for the re-identification of objects lost by STA due to occlusion in many consecutive frames or leaving the FoV. For this reason, lost objects are registered to an archived object list, which is periodically checked by the LTA process [6]. LTA must also recognize when a new, previously not registered person appears in the scene. Finally, we generate a 2D trajectory of each pedestrian. Since the extracted 2D raw object tracks proved to be quite noisy, we applied a 80% compression of the curves in the Fourier descriptor space [172], yielding smoothed tracks (see Fig. 6.13, 6.14).

6.3.3 Lidar based gait analysis

In our approach, the main goal of gait investigation is to support the long-term assignment (LTA) process of the tracking module. To fulfill the requirements of real surveillance systems, we need to extract unique biometric features online during the multi-target tracking process from the measurement sequence.

For gait analysis, we focus on 2D silhouette based approaches [22], which are considered quite robust against low resolution and partial occlusion artifacts, due to capturing information from the whole body. The first step is projecting the 3D points of a person in the RMB Lidar point cloud to an appropriately selected image plane (Fig. 6.13(a)). Since the FoV of the RMB Lidar sensor is circular, a straightforward projection plane could be taken at a given ground position as the local tangent of the circle around the sensor location (see Fig. 6.13(b)). However this choice would not ensure viewpoint invariant features as the silhouette's orientation may be arbitrary. Instead, we interpolate the side view projections of the 3D human silhouettes, by exploiting the assumption that people mostly walk forwards in the scene, turning towards the tangent direction of the trajectory. At each time frame, we project the point cloud segment of each person to the plane, which intersects the actual ground position, is perpendicular to the local ground plane, and it is parallel to the local tangent vector of the Fourier-smoothed trajectory from the top view (Fig. 6.13(a) and (c)).

The projected point cloud consists of a number of separated points in the image plane, which can be transformed into connected 2D foreground regions by morphological operations. A main



(a) Bird's view (b) Circular tangent (c) Trajectory tangent

Figure 6.13: Silhouette projection: (a) a tracked person and its projection plane in the point cloud from bird's view. Variants: (b) the projection plane's normal points towards the sensor (undefined silhouette orientation) (c) the projection plane is the tangent of the trajectory (sideview silhouettes)

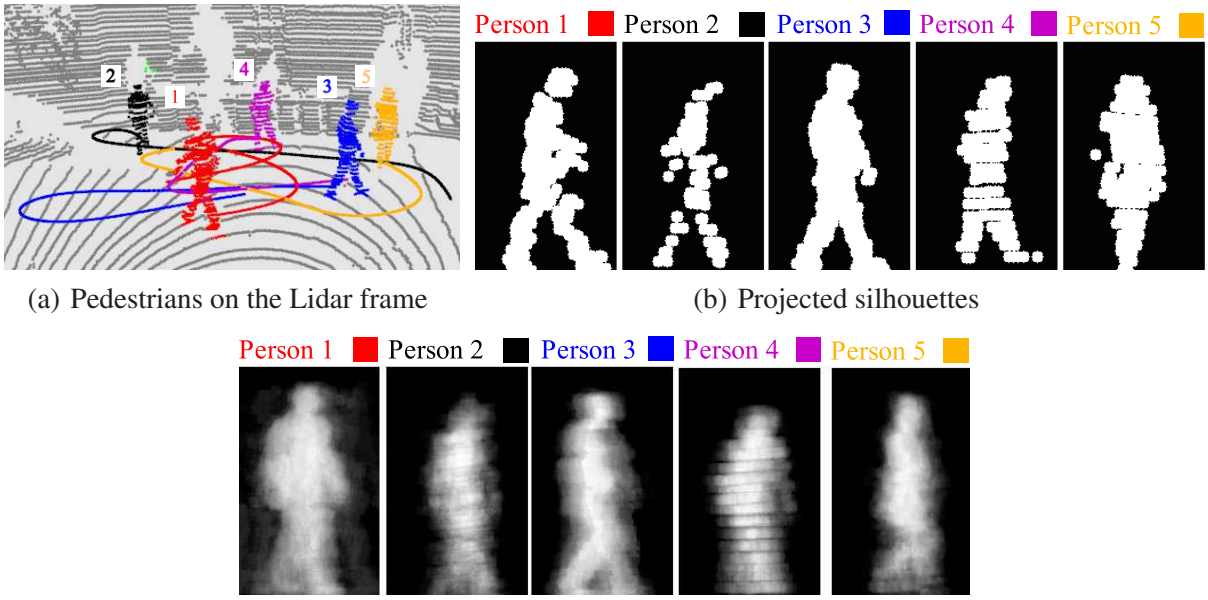
advantage of the Lidar technology is that the laser measurement is directly available in the 3D Euclidean coordinate space, without perspective distortion and scaling effects, thus the projected silhouettes may be also compared without re-scaling. However, the density of the point cloud representing a given person is significantly lower at a larger distance from the sensor, yielding silhouettes which have discontinuities. Challenging samples can be observed in Fig. 6.14(a),(b), which show a snapshot from a 5-person-sequence with the extracted silhouette masks. *First*, the silhouettes of Persons 2 and 4 are disconnected, since they are far away from the sensor. *Second*, for people walking towards the sensor, the 2.5D measurement provides a frontal or back view, where the legs may be partially occluded (see Person 5). *Third*, some silhouette parts may be occluded by other people or field objects in a realistic surveillance scene (see Person 2).

In our proposed model, we adopt the idea of Gait Energy Image (GEI) based person recognition to the Lidar surveillance environment [19, 22]. The original GEI approach was introduced by Han and Bhanu in 2006 [173] for conventional optical video sequences. GEIs are derived by averaging the binary person silhouettes over the gait cycles:

$$G(x,y) = \frac{1}{T}\sum_{t=1}^{T} B_t(x,y) \tag{6.10}$$

where $B_t(x,y) \in \{0,1\}$ is the (binary) silhouette value of pixel (x,y) on time frame t, and $G(x,y) \in [0,1]$ is the (rational) GEI value. In [173] a person was represented by a set of different GEI images corresponding to the different observed gait cycles, which were compressed by Principal Componant Analysis (PCA) and Multiple Discriminant Analysis (MDA). Thereafter person recognition was achieved by comparing the gallery (training) and probe (test) features.

In our environment, a number of key differences had to be implemented compared to the reference model [173], leading to a new descriptor that we call *Lidar-based Gait Energy Image* (LGEI, see Fig. 6.14(c)). The first key contribution is, that since the RMB Lidar measurement sequences



(a) Pedestrians on the Lidar frame (b) Projected silhouettes

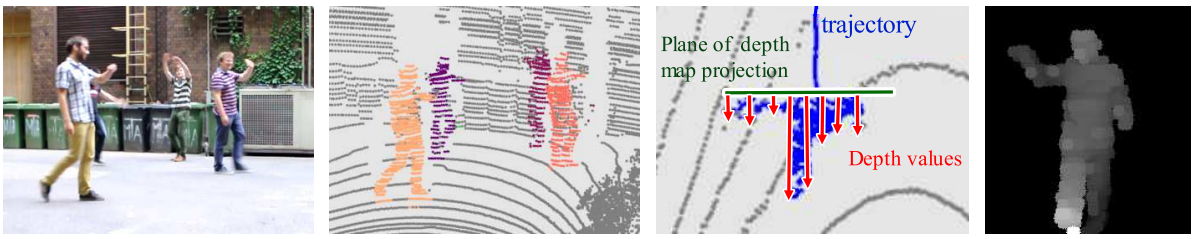
(c) Lidar based Gait Energy Images

Figure 6.14: LGEI generation process: (a) Output of the multi-pedestrian tracker for a sample Lidar frame (person point clouds+trajectories)(b) projected pedestrian silhouettes on the selected Lidar frame (c) Lidar based Gait Energy Images extracted for the people of (b)

have a significantly lower temporal resolution (15 fps), than the standard video flows ($\geq$25 fps), samples from a single gait cycle provide too sparse information. For this reason, we do not separate the individual gait cycles before gait print generation, but we select k (used $k = 100$) random *seed frames* from each person's recorded observation sequence instead, and for each *seed* we average the l consecutive frames (used $l = 60$) to obtain a given LGEI sample. This way, k LGEIs are generated for each individual, and to enable later data compression, global PCA and MDA transforms are calculated for the whole dataset.

The second key difference is, that instead of following the direct GEI-set based person representation and vector comparison of [173], we propose here a neural network based approach. Similarly to [174] we have chosen to use a committee of a Multi-Layer Perceptron (MLP) and a convolutional neural network (CNN), both having N outputs, where N is equal to the number of people in the training scenario. The dominant 35 PCA and 5 MDA components of the LGEIs are used to train an MLP for each person, while the CNN inputs are the raw 2D LGEIs. We used *tanh* activation functions whose output is in the $[-1, 1]$ domain. Thus for a training sample of the ith person, the ith network's prescribed output value is 1, while the remaining outputs are -1.

In the person recognition phase, we generate probe LGEIs for each detected and tracked subject: we start from a random seed frame of the sequence and average the upcoming l consecutive silhouettes. The trained networks produce outputs within the range $o_{\text{MLP}}, o_{\text{CNN}} \in [-1, 1]$, and the



(a) Video ref. frame (b) Lidar frame (c) Projection (d) Depth map

Figure 6.15: (a)-(b): A sample frame from an outdoor test sequence used for activity recognition (c)-(d): Demonstration of the the frontal projection and depth map calculation for activity recognition. Projection plane is perpendicular to the trajectory.

ith output (corresponding to the ith trained person) of the MLP-CNN committee is taken as the maximum of the outputs of the two networks: $o^i = \max(o^i_{\text{MLP}}, o^i_{\text{CNN}})$, $i = 1, \ldots, N$. As a valid identification of a given G probe LGEI, only positive $o^i(G)$ values are accepted. Therefore, with the notation of $i_{\max} = \text{argmax}_i\, o^i(G)$, sample G is recognized as person $i_{\max}$, if $o^{i_{\max}} > 0$, otherwise we mark G as *unrecognized*.

For reducing further artifacts caused by frequent occlusions, we also developed a frame selection algorithm. A binary mask is created by summing and thresholding the consecutive silhouettes for every person. For every silhouette we calculate its internal and external area w.r.t. the mask. If the internal area is less then 40% of the mask's area or the external area is more then 30% of the mask's area the frame is discarded from the LGEI calculation. Note that the above sample collection scheme is is not effected by prior gait cycle estimation in contrast with [173].

6.3.4 Action recognition

The recognition of various actions can provide valuable information in surveillance systems. The main goal of this section is to propose features for recognizing selected - usually rarely occurring - activities in the Lidar surveillance framework, which can be used for generating automatic warnings in case of specific events, and removing various 'non-walk' segments from the training/test data of the gait recognition module.

Apart from normal walk, we have selected five events for recognition: *bend*, check *watch*, *phone* call, *wave* and wave two-handed (*wave2*) actions. A sample outdoor frame with four people is shown in Fig. 6.15. Our approach for action recognition is motivated by the LGEI based gait analysis technique, however, various key differences have been implemented here.

First, while gait could be efficiently analyzed from side-view point cloud projections, the actions listed above are better observable from a frontal point of view. For this reason, we have

(a) walk (b) bend (c) check watch (d) phone call (e) wave (f) wave2

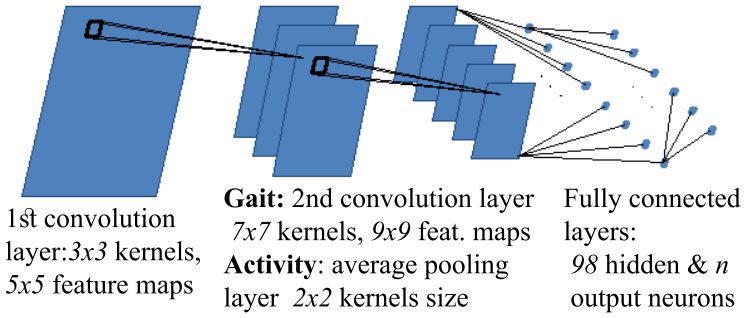
Figure 6.16: ADM (left) and AXOR (right) maps for the different actions.

chosen a projection plane for action recognition, which is perpendicular to the local trajectory tangent, as demonstrated in Fig. 6.15(c).

Second, various actions, such as waving or making phone calls produce characteristic local depth texture-patterns (e.g. the hand goes forward for waving). Therefore, instead of deriving binarized silhouettes, we create depth maps by calculating the point distances from the projection plane according to Fig. 6.15(c), a step which yields a depth image shown in Fig. 6.15(d). Then, we introduce the *averaged depth map* (ADM) feature as a straightforward adoption of the LGEI concept, so that we average the depth maps for the last τ frames, where τ is the a preliminary fixed time window related to the expected duration of the activities (we used $\tau = 40$ frames uniformly). ADM sample images for each activity are shown in Fig. 6.16 (left samples).

Third, while gait is considered a low-frequency periodic motion of the whole body, where we do not lose a significant amount of information by averaging the consecutive images, the above actions are aperiodic and only locally specific for given body parts. For example, waving contains sudden movements, which yield large differences in the upper body regions of the consecutive frames. Thus, apart from ADM we introduce a second feature, called *averaged XOR image* (AXOR), which aims to encode information about the motion dynamics. An exclusive-OR (XOR) operation is applied on two consecutive binarized frontal silhouettes, and the AXOR map is calculated by averaging these binary XOR images and taking the squares of the average values. The AXOR map displays high values for the regions of sudden movements, as shown in Fig. 6.16 (right image of each pair), especially regarding the waving actions in images (e) and (f).

We continue with the description of the training and recognition steps. For each action from the set *bend*, *watch*, *phone*, *wave* and *wave2*, two separate convolutional neural networks (CNN) were trained, one for the ADM and one for the AXOR features, respectively. As explained in [175], a small (4-layer) CNN could be constructed, using the spatially downscaled (to 20×16 pixels) and normalized ADM and AXOR feature maps. During the training of the CNNs, we prescribed the output values 1.0 for positive and -1.0 for negative samples by each activity. The negative training data also included various samples from normal *walking*. The outputs of the CNNs range



Task:	Gait	Activity
k1	3	7
f1	5	5
k2	7	2
f2	9	-
h	98	20
n	NP	1

Figure 6.17: Structure of the used convolutional neural networks (CNN). By gait recognition, N is equal to the number of people in the training set.

from -1.0 to 1.0 , and a probe sample is recognized as a given action if the corresponding ADM-based and AXOR-based CNN outputs both surpass a ν decision threshold (used $\nu = 0.6$). If no activity is detected, we assume that the observed person is in the *walking* state. If multiple CNN outputs surpass the decision threshold, we select the action with the highest confidence.

6.3.5 Dataset for evaluation

Since to our best knowledge no Lidar based gait or activity recognition dataset has been published yet for surveillance environments, we have created the SZTAKI Lidar Gait-and-Activity (SZTAKI-LGA) database[1], which was designed for the evaluation of gait based person identification and activity recognition in a multi-pedestrian environment.

For gait analysis, our proposed SZTAKI-LGA database contains *ten* outdoor sequences captured in a courtyard by a Velodyne HDL-64E RMB Lidar sensor. All the sequences have 15 fps frame rate, their length varies between 79 and 210 seconds (in average 150 sec.), and each one contains 3-8 people walking simultaneously in the scene. In each case, the test subjects were asked to walk naturally in the scene, then all leave the Field of View, re-appear in a different order, and walk till the end of the sequence. This *screen-play* enables to test gait descriptors in realistic surveillance situations, with the goal of matching the corresponding gait patterns collected in the first (*training*) and second (*probe*) parts of each test scenario. Since the sequences were recorded in different seasons, we can also investigate how different clothing styles (such as winter coats or t-shirts) influence the discriminating performance of the observed gait features.

For *action recognition* purposes we recorded 1 indoor and 9 outdoor sequences with a total time of 633 seconds. The test data contains various examples for the five addressed activities: *bend* (88 samples), *watch* (53), *phone* (50), *wave* (58) and *wave2* (46) which are extracted from

[1]Url: `http://mplab.sztaki.hu/geocomp/SZTAKI-LGA-DB`.

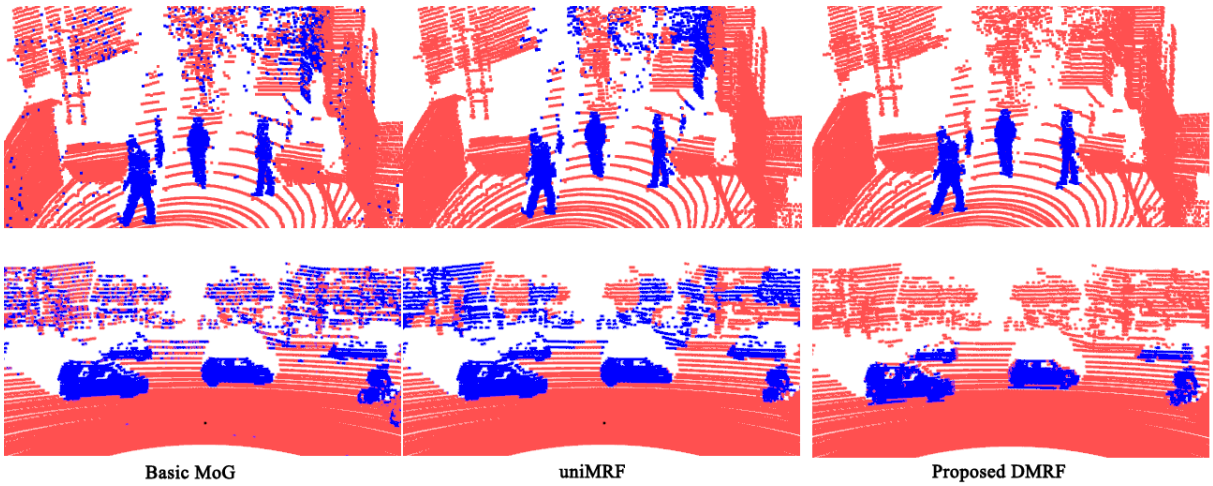


Figure 6.18: Foreground detection results on sample time frames with the *Basic MoG*, *uniMRF* and the *proposed DMRF* models: foreground points are displayed in blue (dark in gray print).

the sequence. Each sequence contains multiple pedestrians, and the typical length range of a given annotated activity sample varies between 40-100 frames.

6.3.6 Experiments and discussion

We have evaluated the proposed algorithms using our recorded real Lidar sequences. The structures of the convolutional neural networks used for gait and activity recognition were similar, only the second layer's type, the number of feature maps and the kernel size parameters were different, as detailed in Fig. 6.17. The MLP component in gait analysis used 6 hidden neurons and N outputs, equal to the number of people in the training scenario.

6.3.6.1 Evaluation of foreground-background separation

We have tested the foreground detection algorithm in various sequences of the SZTAKI-LGA database and also on a *traffic* monitoring scenario (see Fig. 6.18). The *Traffic* sequence was recorded with 5Hz from the top of a car waiting at a traffic light in a crowded crossroad.

We have compared our proposed DMRF model for foreground-background separation to three reference solutions. First, we implemented the *Basic MoG* approach (already introduced in Sec. 6.3.1.2), which is based on [162] with using on-line K-means parameter update [171]. Second, we tested *uniMRF* (detailed in Sec. 6.3.1.3), which partially adopts the uniform foreground model of [160] for range image segmentation in the DMRF framework. Third, we also tested an MRF model in 3D, called *3D-MRF*, based on [176]. We define in *3D-MRF* point neighborhoods in the original $\mathcal{L}^t$ clouds based on Euclidean distance, and use the background fitness values of (6.7) in the data model. The graph-cut algorithm [105] is adopted again for MRF energy optimization.

Table 6.2: Point level evaluation of foreground detection accuracy (F-score in %) and processing speed

Sequence name	**Point cloud size**	F-score based on 100 frames (in %)			
		Bas. MoG	uniMRF	3D-MRF	**DMRF**
`Summer1`	65K pts/fr.	55.7	81.0	88.1	95.1
`Summer2`	86K pts/fr.	59.2	86.9	89.7	93.2
`Summer3`	86K pts/fr.	38.4	83.3	78.7	89.0
`Winter1`	86K pts/fr.	55.0	86.6	84.1	91.9
`Winter2`	86K pts/fr.	54.9	86.6	84.1	91.9
`Spring1`	86K pts/fr.	49.9	84.8	82.7	88.9
`Spring2`	86K pts/fr.	56.8	89.1	86.9	94.4
`Traffic`	260K pts/fr.	70.4	68.3	76.2	74.0
Processing Speed (fps)		120fps	17-18fps	2-7fps	15-16fps

Qualitative segmentation results on sample frames from three sequences are shown in Fig. 6.18, concerning *Basic MoG*, *uniMRF* and the proposed DMRF model. For *quantitative* (numerical) evaluation, we manually generated Ground Truth (GT), through annotation around 100 relevant frames of each test sequence. For quantitative evaluation metric, we have chosen the point level *F-score* of foreground detection. We have also measured the processing speed in frames per seconds (fps). The numerical performance analysis is given in Table 6.2. The results confirm that the proposed model surpasses the reference techniques in *F-score* in all surveillance sequences, meanwhile the processing speed is 15-16fps, which enables real-time operation. In the `Traffic` sequence with large and dense point clouds, the 3D-MRF approach is able to slightly outperform our approach in detection rate, but the *proposed DMRF* method is significantly quicker: we measured there 2fps processing speed with 3D-MRF and 16fps with the proposed DMRF model. We can also observe that differently from 3D-MRF, our range image based technique is less influenced by the size of the point cloud.

6.3.6.2 Evaluation of gait recognition

The gait recognition module has been validated on all the 10 test gait-sequences of the database. For comparison, we implemented three different model-free silhouette or range image based approaches in our Lidar-based surveillance framework, which are Lidar-focused modifications of state-of-the-art methods, proposed earlier for standard optical and Kinect data: *Silhouette Print + Dynamic Time Warping* (SP+DTW) [177] *Depth Gradient Histogram Energy Image* (DGHEI) [178] and the *Color Gait Curvature Image* (CGCI) [179]. Our proposed approach is referred as the *Lidar Based Gait Energy Image* with MLP+CNN committee (LGEI). All the methods (except the silhouette print) were trained using 100 gallery feature maps for each person, extracted from the *training* parts of the sequences. In the evaluation phase, we generated 200 probe maps of each test

subject from the *test* segments of the videos. Each probe sample was independently matched to the trained person models, thus we used $200 \cdot N$ test samples in a scenario with N people. For evaluating the performance of the different methods, we calculated the rate of the correct identifications among all test samples, and listed the obtained results in Table 6.3.

Although according to their introducing publications, both the *CGCI* [179] and *DGHEI* [178] methods proved to be notably efficient for processing Kinect measurements, their advantages could not be exploited by dealing with the much sparser Velodyne RMB Lidar point clouds. In particular, as we can observe in Table 6.3, the *CGCI* method proved to be the less successful among all the tested techniques for the low density Lidar data. By testing the width-vector based *SP+DTW* approach [177], we experienced that it only favored the first test scene (`Winter0`), which included nearly complete silhouettes with noiseless contours. However as the quality of silhouettes decreased due to frequent occlusions, the SP+DTW approach provided quite low recognition rates. The *DGHEI* [178] proved to be the second best gait descriptor, outperformed only by our LGEI based method by 5% overall. Note that DGHEI has originally been proposed by extending the Gait Energy image (GEI) with depth gradient extraction and direction histogram aggregation, which improvement increased the performance when high-quality depth images were available. However, in our scenarios with lower resolution depth maps these features could not be efficiently utilized.

Our proposed LGEI solution has been tested first by separately using the MLP and CNN networks, and thereafter with the MLP+CNN committee. As the last three columns of Table 6.3 confirm, the MLP and CNN outperformed each others on a case-by-case basis, and the committee has generally resulted in improved results over the two network components. As already shown in [21], in LGEI classification the MLP-CNN committee could also outperform the standard Vector Comparison approach proposed in [173].

Table 6.3 also demonstrates that compared to the SP+DTW, DGHEI and CGCI techniques the LGEI method provided superior results in most of the test scenarios. The performance drop observed by some of the *more crowded* (6-8 person) scenes has been principally caused by the increased number of occlusions which obviously yielded lower quality input data for the classification framework. As examples, the score matrices between the trained neural networks and the measured gait patterns from the different test subject are displayed in Fig. 6.19 for five test scenes. This figure highlights the background of the varying performance in the different test cases: from the point of view of (LGEI-based) gait recognition `Spring0` and `Summer0` proved to be *simple* scenarios with nearly *diagonal* score matrices, while `Spring1` and `Summer1` are quite *difficult* sequences, where the measurable benefits of the LGEI technique are the most apparent compared to the weaker performing reference approaches.

By further examination of the LGEI method, we investigated the improvements caused by two auxiliary innovations of our proposed approach: First, applying trajectory tangent (TT) oriented

Table 6.3: Evaluation of gait based person re-identification with different methods.

Scene	Num. of people	SP+ DTW	DG-HEI	CGCI	LGEI		
					CNN	MLP	Mix
Winter0	4	0.96	0.97	0.36	0.94	0.98	**0.99**
Winter1	6	0.33	0.89	0.27	0.85	0.90	**0.95**
Spring0	6	0.64	0.81	0.32	0.91	0.95	**0.98**
Spring1	8	0.33	0.59	0.20	0.63	0.66	**0.70**
Summer0	5	0.39	0.97	0.40	0.99	0.95	**1.00**
Summer1	6	0.33	0.83	0.29	0.77	0.95	**0.95**
Summer2	3	0.33	0.98	0.53	0.96	0.99	**0.99**
Summer3	4	0.50	**0.94**	0.32	0.94	0.93	**0.94**
Summer4	4	0.25	**0.95**	0.27	0.91	0.90	0.91
Summer5	4	0.50	**0.80**	0.32	0.77	0.74	**0.80**
Average	5	0.46	0.87	0.33	0.87	0.90	**0.92**

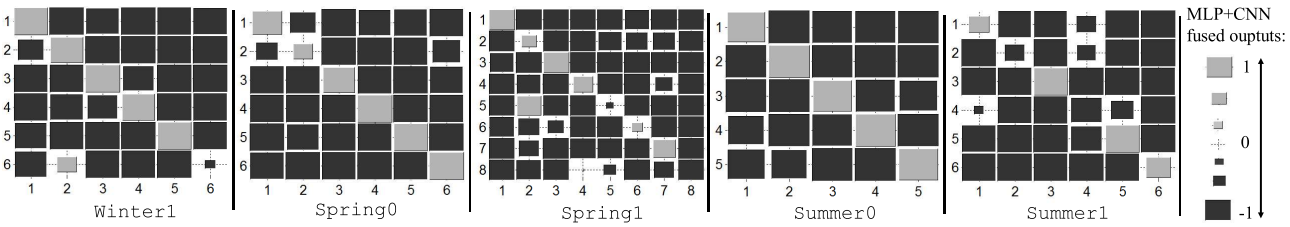
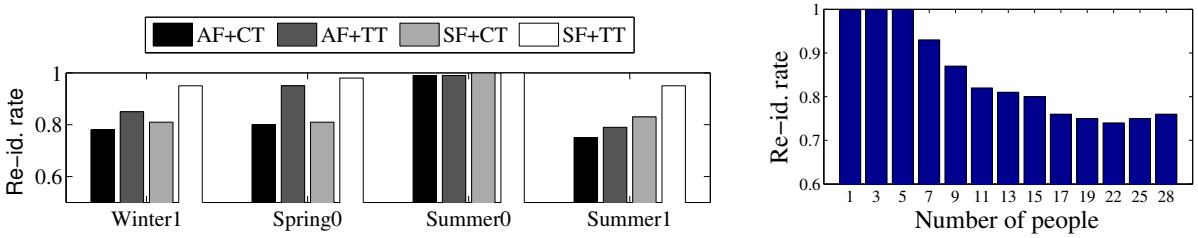


Figure 6.19: Quantitative evaluation of LGEI based matching between the gallery (columns) and probe (rows) samples. Rectangles demonstrate the CNN+MLP outputs, the Ground Truth match is displayed in the main diagonal.



(a) Result of CT and TT improvements

(b) Results on the whole dataset

Figure 6.20: (a) Performance improvements caused by trajectory based projection plane estimation (TT) over circular tangent (CT), and frame selection (SF) over all frames (AF) strategy using the LGEI method. (b) Performance as a function of number of people

Table 6.4: The confusion matrix, and precision/recall rates of action recognition for each event.

Detect→ / GT↓	*Bend*	*Watch*	*Phone*	*Wave*	*Wave2*	*All*	*FN*	*FP*	*Pr*	*Rc*
Bend	85					88	3		1.00	0.97
Watch		37	1		4	53	11	3	0.82	0.77
Phone		5	36	2	2	50	5	6	0.69	0.88
Wave			4	44	5	58	5	3	0.76	0.90
Wave2			5	9	31	46	1	2	0.70	0.97

planes of silhouette projection instead of the straightforward circular tangent (CT) direction (refer to Fig. 6.13). Second, automatic selection of frames (SF) instead of using all frames (AF) in LGEI generation, by dropping the presumptively low quality silhouettes (Sec. 6.3.3). As shown in Fig. 6.20(a) for four selected sequences, both new algorithmic steps yielded notable improvements in the recognition rates.

Our next evaluation stage addresses the performance variation of LGEI based gait recognition, by increasing the number of people in the database. Exploiting that in our 10 test sequences 28 different people have appeared, we collected the silhouette sequences of the different test subjects from all test scenarios into a global database. Then, we selected step by step 2,3,...,28 people from the database, and each time we trained and evaluated an LGEI-CNN+MLP committee for the actual subset of the people (using separated training and test samples). The diagram of the observed recognition rates as a function of the number of persons is displayed in Fig. 6.20(b), which shows a graceful degradation in performance, staying steadily around 75% for 17-28 people.

We also compared the computational time requirements of the different approaches [1] and we have experienced the the proposed LGEI technique is also competitive at this point. Although it needs relatively significant time for training set generation and training of the CNN and MLP networks, but the recognition step is still very efficient: less then 0.01sec/probe sample.

6.3.6.3 Evaluation of activity recognition

For evaluating the proposed activity recognition module, we used the ten activity sequences of the database, applying a cross validation approach. For testing the recognition performance on each sequence, we trained the actual CNNs with the manually annotated activity patterns of the other nine sequences. For both training and recognition we also used various negative samples cut from normal walking parts of the scenarios. The number of selected *walk* frames was equal to the average number of frames corresponding to the other activities.

For presenting the result, the aggregated confusion matrix of action recognition in the test scenes is shown in Table 6.4. The matrix value of the ith row and jth column indicates the num-

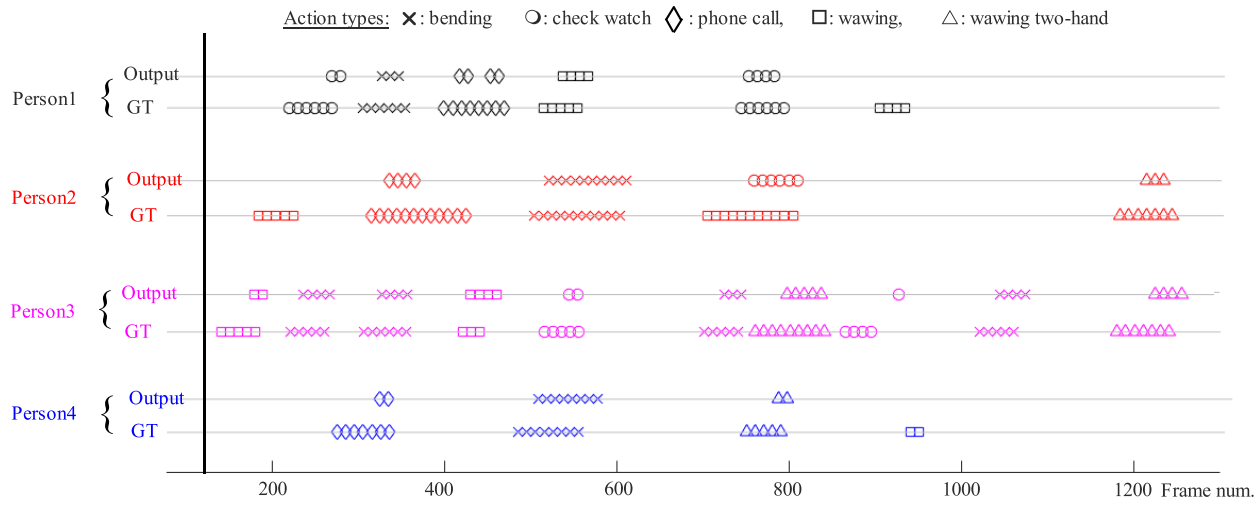


Figure 6.21: Result of *activity recognition* in an outdoor test sequence (4 people).

ber of samples from the ith activity, which were recognized as action j. The last two columns correspond to false negative (FN) and false positive (FP) detections: for row i, FN is equal to the number of ignored occurrences of the ith action, which were neither identified by any of the other activities, while FP is the number of erroneous alerts of the ith activity in the case when none of the addressed events occurred.

As we can see, the *bend*, *phone*, *wave* and two-handed waving (*wave2*) activities were almost always denoted as an event (FN≤ 5), while check *watch* indicated 11 false negative samples, since the small hand movements were occasionally imperceptible due to occlusions. *Bend* was never confused with other actions, while *wave* and *wave2* were mixed up in a number of cases. It is also worth noting that the overall number of false positives is quite low (Σ_iFP$< 5\%$ of the real events), i.e. the system rarely indicates unexpected warnings in case of normal walks. This advantageous property can be well examined in the timeline diagram displayed in Fig. 6.21, which corresponds to one of the outdoor test sequences. The horizontal axis corresponds to the frame index, and the different activities are denoted by different markers (as explained in the top row). For each person, the *Output* row marks the frames where our approach detected various activities, while the *GT* (Ground Truth) row indicate the manually annotated reference frames. In agreement with Table 6.4, in nearly all cases the real activities are detected by the system with a time delay necessary for ADM and AXOR map generation.

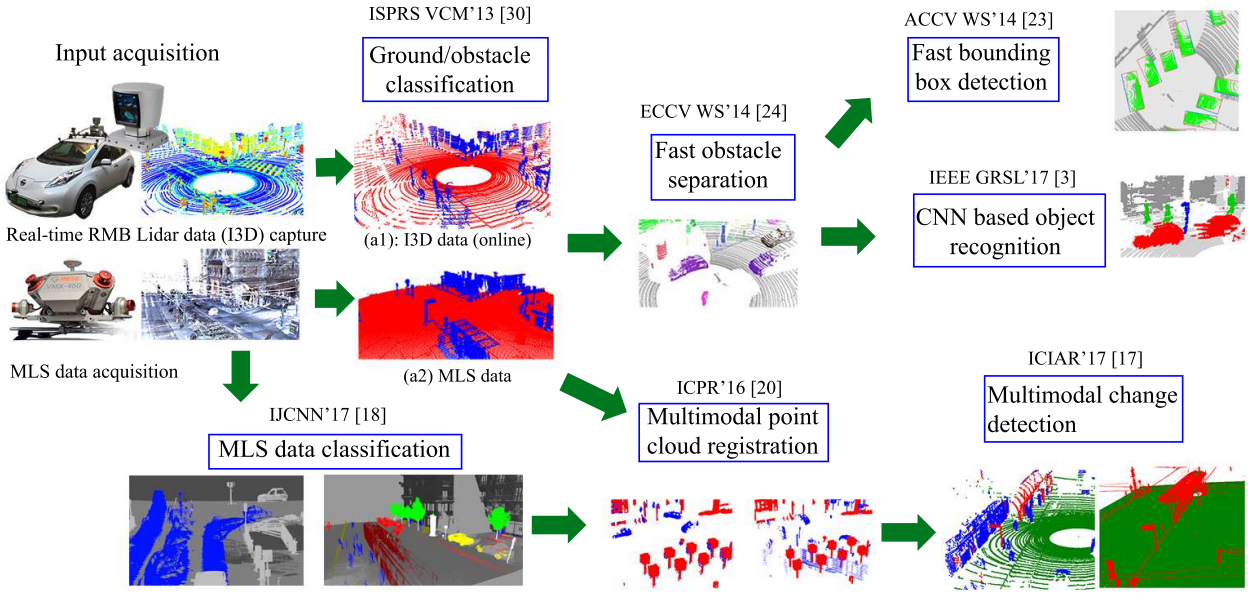


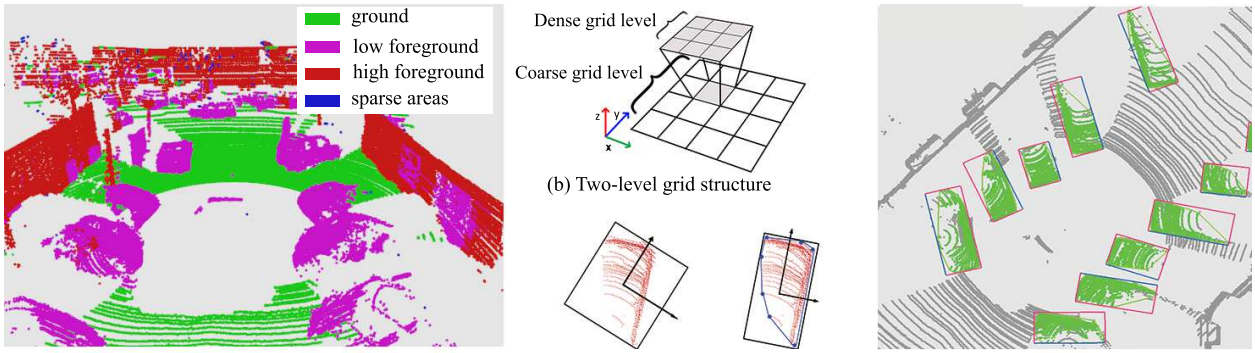
Figure 6.22: Workflow of instant environment perception, composed of the proposed algoritms. Publications serving as basis of this section are noted above the algorithm boxes.

6.4 Urban scene analysis with real time Lidar sensors and dense MLS data background

This section introduces various new algorithms related to real time perception of dynamic urban environments with a car-mounted RMB Lidar scanner. As high resolution 3D background map, we utilize dense point clouds obtained by mobile laser scanning (MLS).

The workflow of the conducted research work is briefly summarized in Fig. 6.22. The first four steps work solely on the RMB Lidar streams. The process starts with *ground-obstacle separation*, which must be performed in real time and robustly, dealing also with lower quality (non-planar) road surfaces often appearing in minor roads of cities [29, 30]. Within the obstacle class, we distinguish *low foreground* and *high foreground* regions, which will help the subsequent steps. Then, we perform *fast object separation* in the foreground areas of the sparse and inhomogeneous RMB Lidar point clouds [24], while a quick bounding box estimation algorithm will also be presented to support the analysis of field objects [23]. Thereafter, the separated object blobs in the *low foreground* are classified via a deep neural network into vehicle, pedestrian, street furniture, and wall component classes, with the help of anchor objects from the *high foreground* [3, 26].

The remaining steps in Fig. 6.22 aim at the efficient joint utilization of the dynamic RMB Lidar based and static MLS measurements. First we perform a 3D CNN-based semantic classification of the dense MLS point clouds [18], which enables filtering out all moving and movable objects from



(a) Segmented Velodyne frame (b) Two-level grid structure (c) PCA vs. proposed bounding box fitting (d) Detected objects and bounding boxes

Figure 6.23: RMB Lidar data segmentation and object detection

the high resolution background map. Then, using a coarse Global Positioning System (GPS) based initial positioning of the vehicle, we propose a fast and accurate registration algorithm between sparse RMB Lidar frames and the dense MLS point clouds [20], which facilitates centimeter-accuracy localization of the vehicle in the HD map [16]. Finally, we construct a *multimodal change detection* approach [17] to identify regions of the actual RMB Lidar point cloud frames, which are not present in the MLS background model.

6.4.1 Ground-obstacle classification

In this step, the input point cloud is segmented into four regions: *ground*, *low foreground*, *high foreground* and *sparse areas*. By our definition, low foreground is the estimated region of short street objects, such as cars, pedestrians, benches, mail boxes, billboards etc, while high foreground covers tall objects, among others building walls, trees, traffic signs and lamp posts.

Point cloud segmentation is achieved by a grid based approach [30]. We fit a regular 2D grid S with fixed rectangle side length onto the $P_{z=0}$ plane (using the RMB Lidar sensor's vertical axis as the z direction), where $s \in S$ denotes a single cell. We assign each $p \in \mathcal{P}$ point of the point cloud to the corresponding cell s_p, which contains the projection of p to $P_{z=0}$.

We use point height information for assigning each grid cell to the corresponding cell class. Before that, we detect and remove *sparse* grid cells which contains less points than a predefined threshold (used 8 points). After clutter removal all the points in a cell are classified as *ground*, if the difference of the minimal and maximal point elevations in the cell is smaller than a threshold (used $25cm$), and the average elevation in the neighboring cells does not exceed an allowed height range. A cell belongs to the class *high foreground*, if either the maximal point height within the cell is larger than a predefined value (used $140cm$ above the car top), or the observed point height difference is larger than a threshold (used $310cm$). The rest of the points in the cloud are assigned to class *low foreground*. A segmented frame is shown in Fig. 6.23(a).

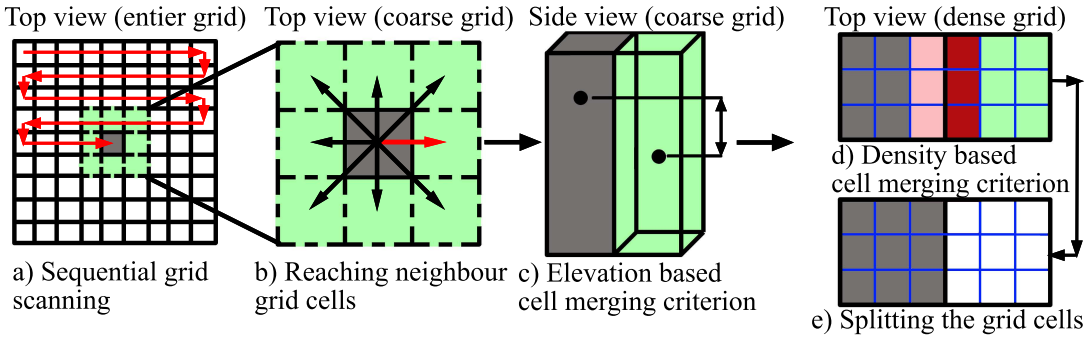


Figure 6.24: The step by step demonstration of the object detection algorithm

Due to the limited vertical view angle of the RMB Lidars (for Velodyne HDL-64E: +2° up to -24.8° down), the defined elevation criteria may fail near to the sensor position. In narrow streets where road sides located closely to the measurement position, several nearby grid cells can be misclassified regularly *e.g.* some parts of the walls and the building facades are classified to *low foreground* cell class instead of *high foreground* cell class (see in Fig. 6.23(a)). By definition, we will refer to these misclassified wall segments henceforward as *short facades*, which should be detected and filtered out at a later step by the object detector.

6.4.2 Fast object separation and bounding box estimation

After the point cloud segmentation step, our aim is to find distinct groups of points which belong to different urban objects within the *low* and *high foreground* regions, respectively. For this task we introduced a hierarchical grid model [24] (see Fig. 6.23(b)): On one hand, the coarse grid resolution is appropriate for a rough estimation of the 3D blobs in the scene, thus we can roughly estimate the size and location of the possible object candidates. On the other hand, using a dense grid resolution, it is efficient to calculate point cloud features from smaller subvolumes of the scene, therefore we can refine the detection result derived from the coarse grid resolution.

The following two-level grid based connected component algorithm is separately applied for the sets of grid cells labeled as short and tall objects, respectively. *First*, we visit every cell of the coarse grid and for each cell s we consider the cells in its 3×3 neighborhood (see Fig. 6.24(a),(b)). We visit the neighbor cells one after the other in order to calculate two different point cloud features: (i) the maximal elevation value $Z_{max}(s)$ within a coarse grid cell and (ii) the point cardinality in a dense grid cell. *Second*, our intention is to find connected 3D blobs within the foreground regions, by merging the coarse level grid cells together. We use an elevation-based cell merging criterion to perform this step. $\psi(s, s_r) = |Z_{max}(s) - Z_{max}(s_r)|$ is a merging indicator, which measures the difference between the maximal point elevation within cell s and its neighboring cell s_r. If the ψ indicator is smaller than a predefined value, we assume that s and s_r belong to the

same 3D object (see Fig. 6.24(c)). *Third*, we perform a detection refinement step on the dense grid level (Fig. 6.24(c),(d)). The elevation based cell merging criterion on the coarse grid level often yields that nearby and self-occluded objects are merged into a same blob. We handle this issue by measuring the point density in each sub-cell s'_d at the dense grid level. A super-cell is divided into different parts, if we find a separator line composed of low density sub-cells at the fine resolution. Experiments [24] confirm that using this approach, nearby objects, which were erroneously merged at the coarse level, could be often appropriately separated at the fine level.

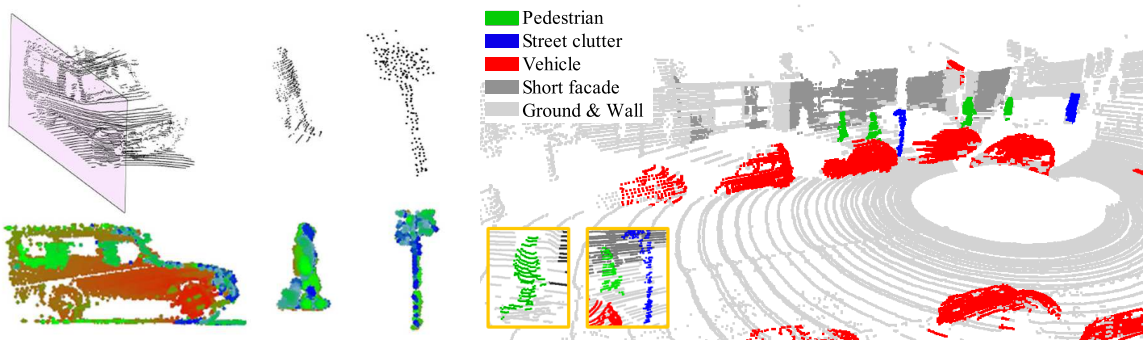
We also proposed a fast 2D bounding box fitting algorithm for cluttered and partially incomplete objects [23]. It is highly challenging to fit precise bounding boxes around the objects in RMB Lidar range data streams, since we should expect various artifacts of self-occlusion, occlusion by other objects, measurement noise, inhomogeneous point density and mirroring effects. These factors drastically change the appearances of the objects, and the conventional principal component analysis (PCA) based techniques [180, 181] may not give sufficient results (see Fig. 6.23(c)). Therefore we calculate the 2D convex hull of the top-view projection of the objects, and we derive the 2D bounding boxes directly from the convex hull (the algorithm is detailed in Appendix E). As shown in Fig. 6.23(c) and (d), this strategy is less sensitive to the inhomogeneous point density and the presence of missing/occluded object segments, since instead of calculating spatial point distributions for the entire object's point set, we capture here the local shape characteristics of the visible object parts, and fit appropriate 2D bounding boxes with partial matching.

6.4.3 Deep learning based object recognition in the RMB Lidar data

Our next main goal is to identify the vehicle and pedestrian objects among the set of connected point cloud segments extracted in Sec. 6.4.2. Our general assumption is that the focused two object classes are part of the low foreground regions, therefore we start with an appearance based classification of the previously obtained short object candidates.

Our labeling considers four object classes. Apart from the *vehicle* and *pedestrian* classes, we create a separate label for the *short facades*, which appear in the low foreground due to the limitations of the height measurement. The remaining short street objects (benches, short columns, bushes etc) are categorized as *street clutter*. Object recognition is performed in a supervised approach: 2D range images are derived from the object candidates, which are classified by a deep neural network. The classification output for each input point cloud sample consists of four confidence values estimating the class membership probabilities for vehicles, pedestrians, short facades and street clutter, respectively.

To obtain the feature maps, we convert the object point clouds into regularly sampled depth images, using a similar principle to [182], but with implementing a number of differences. *First*, we attempt to ensure side-view projections of the objects, by estimating the longitudinal cross



(a) Range image formation (b) Object classification

Figure 6.25: Object classification workflow for RMB Lidar frames

section of the object shapes, using the bounding box estimation algorithm from Sec. 6.4.2. *Second*, we calculate the distance between the estimated plane and the points of the object candidate, which can be interpreted here as a depth value. In order the avoid occlusions between overlapping regions *i.e.* multiple 3D point projections into a same pixel of an image plane with different depth values, we sort the depth values in an ascending order, and we project them to the image plane starting from the closest to the farthest. As demonstrated in Fig. 6.25(a) this projection strategy ensures that object points in the front side do not become occluded by the object points in the back.

For object recognition, we trained a *Convolutional Neural Network (CNN)* based feature learning framework called *Theano* firstly introduced by [183]. The CNN framework receives the previously extracted depth images as an input layer scaled for the size of 96×96, and the outputs are four confidence values from the [0,1] range, describing the fitness of match to the four considered classes: vehicle, pedestrian, short facade and street clutter. In this way, we can later utilize not only the index of the winner class, but also describe how sure the CNN module was about its decision for a given test sample. After testing various different layer configurations, we experienced that four pairs of convolution-pooling layers followed by a fully connected dense layer give us the most efficient results. Finally, in post processing, we extended our approach with a contextual refinement step, exploiting topological constraints between various scene objects using specific *high foreground* objects, called *anchor facades*, as landmarks (details are provided in [3]).

6.4.4 Semantic MLS point cloud classification with a 3D CNN model

We have proposed a new 3D CNN based semantic point cloud segmentation approach, which is adopted to dense point clouds of large scale urban environments, assuming the presence of high variety of objects, with strong and diverse *phantom* effects caused by scene objects moving concurrently with the MLS platform [18]. Our technique is based on a sparse voxel based representation

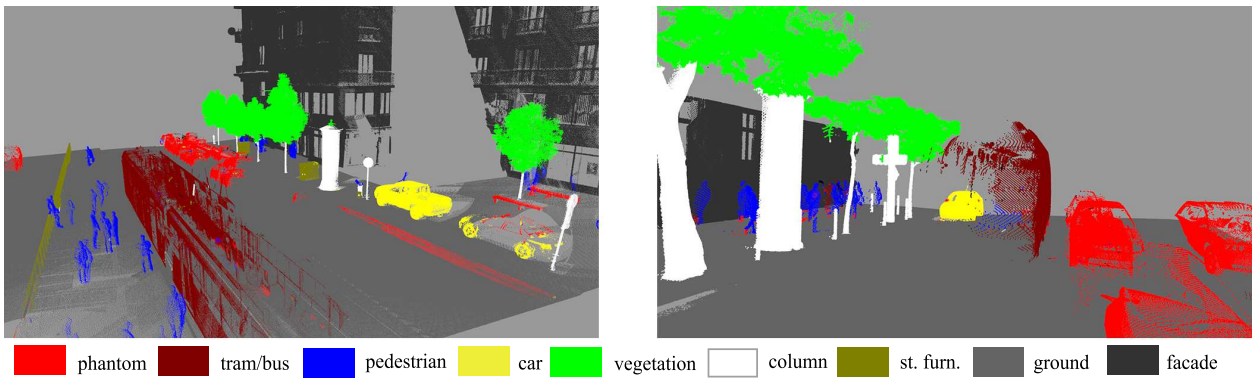


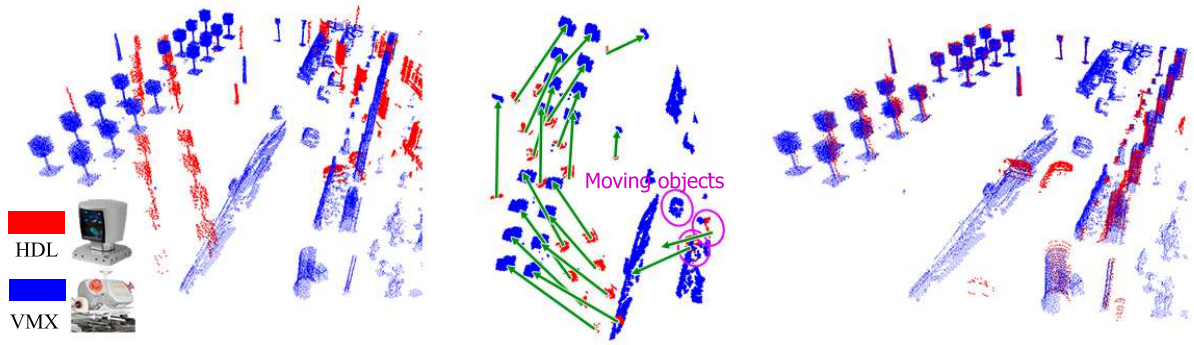
Figure 6.26: Point cloud segmentation result with a 3D CNN

of the scene (with fine 0.1m voxel resolution), and classifies each voxel into one of the following nine semantic classes: *phantom*, *tram/bus*, *pedestrian*, *car*, *vegetation*, *column*, *street furniture*, *ground* and *facade*.

During the data mapping, we assign two feature channels to the voxels based on the input cloud: point *density*, taken as the number of included points, and *mean elevation*, calculated as the average of the point height values in the voxel. The unit of training and recognition in our network is a $K \times K \times K$ voxel neigborhood (used $K = 23$), called hereafter training volume. To classify each voxel v, we consider the point density and elevation features in all voxels in the v-centered training volume, thus a given voxel is labeled based on a 2-channel 3D array derived from K^3 local voxels. Our 3D CNN network contains a feature extractor part using a combination of several 3D convolution, max-pooling and dropout layers, and a second part with fully connected dense layers, which learn the different class models. To segment a scene, we move a sliding volume across the voxelized input point cloud, and capture the $K \times K \times K$ neighborhood around each voxel. Each neighborhood volume is separately taken as input by the two channel CNN classifier, which predicts a label for the central voxel only. We have validated the efficiency of the approach in diverse and real test data from various urban environments, sample results are shown in Fig. 6.26.

6.4.5 Multimodal point cloud registration

We have proposed a solution [20] to robustly register the sparse point clouds of the RMB Lidar sensor mounted on a moving platform to the dense MLS point cloud data, starting from a GPS based initial position estimation of the vehicle (see Fig. 6.27). Although we can find in the bibliography widely used general point cloud registration techniques, such as variants of the Iterative Closest Point (ICP) [184], and the Normal Distribution Transform (NDT) [185], they all need as initial condition a sufficiently accurate preliminary alignment between the input point clouds. Expecting



(a) Pre-allingment of the point clouds (b) Object based scan matching (c) Result of point cloud registration

Figure 6.27: Velodyne HDL-64E to Riegl VMX-450 point cloud registration results

that in our application field significant translation (up to 10m) and large orientation difference must be compensated by the registration method, we have constructed a two-step approach: first, we estimate a coarse transform between the point cloud frames at object level, which step is followed by accurate registration refinement using the standard NDT algorithm.

Our process begins with abstract object extraction both in the RMB Lidar frame and in the MLS point cloud segment, using the fast object separation algorithm introduced in Sec. 6.4.1, which provides two sets of object centers *C1* and *C2*. Similarly to the fingerprint minutia matching approach by [186], we estimate the optiomal transformation parameters between *C1* and *C2* using the generalized Hough transform: we discretize the set of all allowed transformations, then for each transformation we calculate a matching *fitness* score. Finally the transformation with the highest score is taken as result.

Since the Lidar point clouds reflect the true object distances from the 3D world, we can consider the transformation as a composition of translation and rotation only. Note as well that since the vehicles carrying the sensors are moving on urban roads, which rarely contain sudden steep slopes, orientation difference is mainly expected around the vertical z axis of the captured point cloud's local coordinate system, while translation in the x and y direction, along the $P_{z=0}$ horizontal plane. Exploiting that this object level step only aims to find an approximate solution for the matching, we project the point clouds to their $P_{z=0}$ plane, and estimate the 2D translation and scalar rotation in this image plane, as demonstrated in Fig. 6.27(b). In this way, the searched transformation takes the following form:

$$\mathfrak{T}_{dx,dy,\alpha}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} dx \\ dy \end{bmatrix}$$

The space of transformation consists of triplets (dx, dy, α), where each parameter is discretized into a finite set of values.

Fitness scores for the transformation candidates are collected in the accumulator array A, where the $A[dx, dy, \alpha]$ element counts the evidence for the concerning $\mathfrak{T}_{dx,dy,\alpha}$ transformation. The A array can be filled in an iterative way. For each object pair $(o1, o2)$ where $o1 = (x_1, y_1)$ is a point in the set *C1* and $o2$ is a point in the set *C2* we determine all possible $\mathfrak{T}_{dx,dy,\alpha}$ transformations that map $o1$ to $o2$ and we increment the evidence for these transformations in the array. Here we exploit that for every possible rotation value α there is a unique translation vector $[dx, dy]^\top$ such that $\mathfrak{T}_{dx,dy,\alpha}(o1) = o2$, and it can be calculated as:

$$\begin{bmatrix} dx \\ dy \end{bmatrix} = o2 - \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix} o1$$

The obtained dx and dy values need to be quantized to the nearest bins for appointing the actually increaseable element of the A array. The complete pseudo code of the scan alignment method is shown in Appendix E (Algorithm E.2).

Although the above object based scan matching process proved to be largely robust for the considered urban point cloud scenes, its accuracy is limited by the considered planar translation and rotation transformation constraints, and the inaccuracies of object center estimation from the different point clouds. As we detailed in [30], due to the special data acquisition technology used in mobile laser scanning, the ground-less *obstacle* cloud can be efficiently used for automated scene matching with the Normal Distribution Transform (NDT) [185] in case of a high quality initial transformation estimation, which is available in our case by taking the output of the object-level step. Therefore in the proposed registration approach, we transform first the *obstacles* cloud according to the obtained optimal $\mathfrak{T}_{dx,dy,\alpha}$, thereafter we apply NDT for the resulting clouds (see line 16 of Algorithm E.2).

6.4.6 Frame level cross-modal change detection

The *change detection* module receives a co-registered pair of RMB Lidar and and MLS point clouds. Since the MLS data acts here as a detailed 3D background model, as shown in Fig. 6.28, we eliminate all moving our movable elements (such as pedestrians, vehicles, phantoms) from the MLS point cloud using the semantic labeling module of Sec. 6.4.4.

Our proposed solution extracts changes in the range image domain. The range image $I_{\text{RMB}} : \{d_s | s \in S\}$ over a 64×1024 pixel lattice S is generated from the RMB Lidar's point stream in the same way as in Sec. 6.3.1.1, where missing pixel values are interpolated from their 8-neighborhood. The reference background range image $I_{\text{MLS}} : \{a_s | s \in S\}$ is generated from the 3D MLS point cloud with ray tracing, exploiting that the current position and orientation of the RMB Lidar platform are available in the reference coordinate system as a result of the point cloud

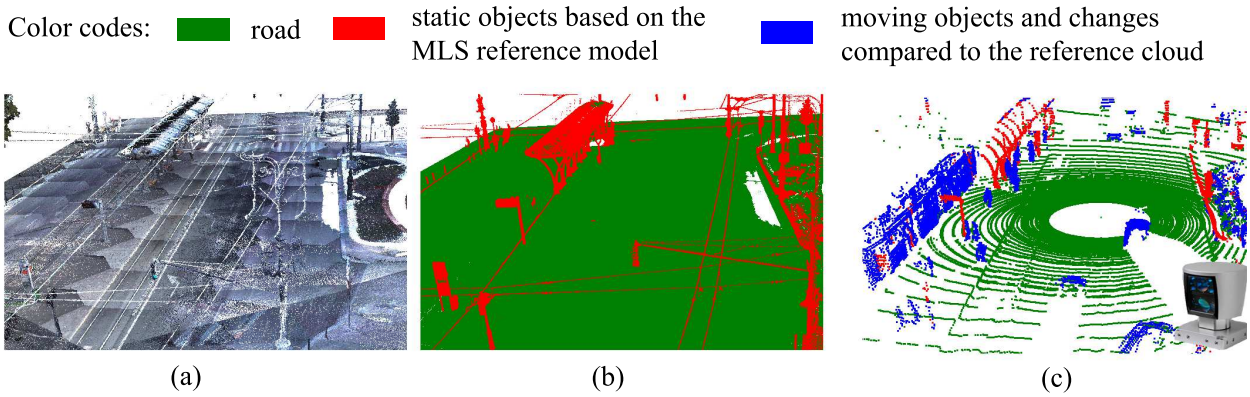


Figure 6.28: Change detection between reference MLS data (a,b) and instant RMB Lidar frames (c)

registration step. Thereafter simulated rays are emitted into the MLS cloud from the moving platform's center position with the same vertical and horizontal resolution as the RMB Lidar scanner. To handle minor registration issues and sensor noise, each range image pixel value is interpolated from MLS points lying inside a pyramid around the simulated RMB Lidar ray [17]. A sample range image pair generated by the above process is shown in Fig. 6.29(a) and (b).

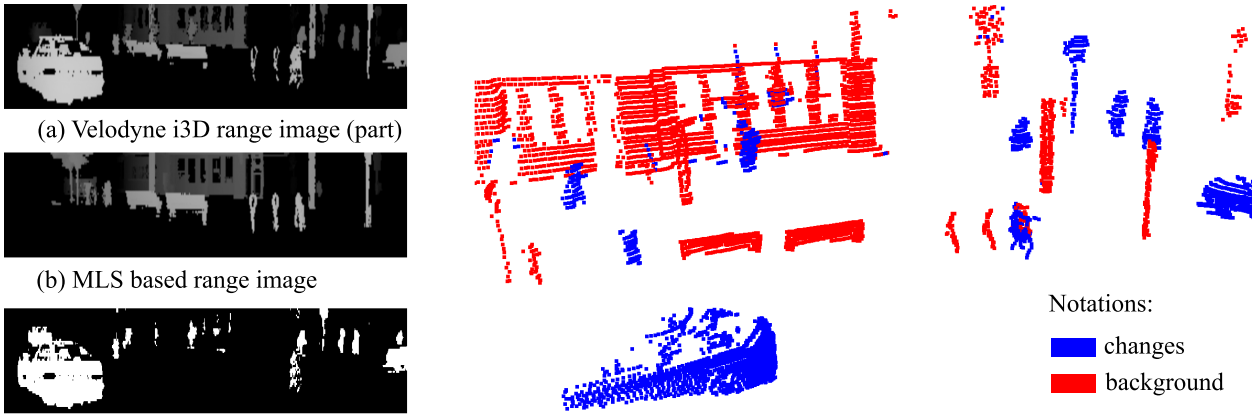
In the next step, the calculated RMB Lidar-based I_{RMB}, and MLS-based I_{MLS} range images are compared using a Markov Random Field (MRF) model, which classifies each pixel of the range image lattice s as foreground (fg) or background (fg). Foreground pixels represent either moving/mobile objects in the RMB Lidar scan, or various environmental changes appeared since the capturing date of the MLS point cloud. Following the notations from eq. (6.9), to formally define the range image segmentation task, we assign to each s pixel of the pixel lattice S a $\varsigma_s \in \{\text{fg}, \text{bg}\}$ class label so that we aim to minimize the following energy function:

$$E = \sum_{s \in S} V_D(d_s, a_s|\varsigma_s) + \sum_{s \in S} \sum_{r \in \mathcal{N}_s} \beta \cdot \mathbf{1}\{\varsigma_s \neq \varsigma_r\}, \tag{6.11}$$

where we used a $\beta = 0.5$ smoothness parameter. The data terms are derived as :

$$V_D(d_s, a_s|\text{bg}) = -\log\left(1 - \frac{1}{1 + e^{d_s - a_s}}\right), \quad V_D(d_s, a_s|\text{fg}) = -\log\left(1 - \frac{1}{1 + e^{-(d_s - a_s)}}\right)$$

The MRF energy (6.11) is minimized via the fast graph-cut based optimization algorithm [105], which process results in a binary change mask in the range image domain, as shown in Fig. 6.29(c). The final step is *label backprojection* from the range image to the 3D point cloud (see Fig. 6.29(d)), which can be performed in a straightforward manner, since in our I_{RMB} range image formation process, each pixel represents only one RMB Lidar point.



(a) Velodyne i3D range image (part)

(b) MLS based range image

(c) MRF based change mask

(d) Backprojection of the change mask to the Velodyne point cloud

Figure 6.29: Demonstration of the proposed MRF based change detection process in the range image domain, and result of label backprojection to the 3D point cloud

6.4.7 Evaluation

The new algorithms and methods proposed in this section have been evaluated versus state-of-the art solutions one by one.

For testing the improvements in object separation and classification (Sec. 6.4.1-6.4.3), we created a new hand labeled dataset, called `SZTAKI Velo64Road`, based mainly on point cloud sequences recorded by our car-mounted Velodyne HDL-64E Lidar scanner in the streets of Budapest[1]. First we run the segmentation (Sec. 6.4.1) and object extraction (Sec. 6.4.2) steps of our model on the raw data, thereafter we annotated all the automatically extracted short objects (2063 objects alltogether) without any further modification with the labels *vehicle*, *pedestrian*, *short facade* and *street clutter*. To demonstrate that the training results are suitable for various urban scenes, we have also validated the performance of our trained model in the `Washington` dataset [187].

First, the object separation module (Sec. 6.4.2) was evaluated, by comparing the automatically extracted object blobs to a manually labeled Ground Truth configuration. As described in [24], we counted the true positives, missing objects and false objects, thereafter we calculated the *F-score* of the detection at object level. As reference of the proposed 2-level grid based model we used a 3D connected component analysis (3D-CCA) approach implemented in [188]. While in *F-score* the proposed approach presented a 13% performance gain versus CCA (84% vs. 71%), meantime it decreased the running speed by two orders of magnitude due to eliminating the kd-tree building step at each frame (27 fps vs. 0.30 fps in average, measured over 1800 sample frames), details can be found in Appendix E, in Table E.1.

[1]Url: `http://mplab.sztaki.hu/geocomp/SZTAKI-Velo64Road-DB.html`

(a) Object classification

Object cat.	Correspond. grouping [188]			Proposed method [3]		
	Pr	Rc	Fs	Pr	Rc	Fs
Vehicle	71	84	77	98	99	99
Short facade	79	52	62	93	77	84
Street clutter	87	93	90	92	97	94
Pedestrian	66	57	61	78	78	78
Overall	76	72	74	90	87	89

(b) Change detection in sidewalk areas

Scene	Voxel based method [191]			Proposed method [17]		
	Pr	Rc	Fs	Pr	Rc	Fs
Deák	81	71	76	87	89	88
Astoria	88	81	84	84	100	91
Kálvin	89	96	92	98	87	99
Fővám	84	64	73	81	97	88
Overall	86	78	81	90	93	92

Table 6.5: Evaluation of the object classification and change detection steps. Notations: Precision (Pr), Recall (Rc), F-score (Fs), in %

Next, we examined the object classification step (Sec. 6.4.3), by evaluating the appearance based labeling, and also the context based refinement. For training the CNN classifier we separated 904 objects from our dataset, which was completed with 434 selected samples Sydney Urban Object Dataset [182]. The test data was composed of the remaining 1159 objects of the `SZTAKI Velo64` dataset. During the evaluation object level *precision*, *recall* and *F-score* values of the detection for each class separately and cumulatively as well. As an independent reference technique, we considered a similar object matching algorithm to [189] based on a corresponding grouping (CG) technique [188]. Comparative results (Table 6.5(a)) showed a 89% overall performance of the proposed approach, and a 15% gain versus the CG reference, while the contextual refinement caused around 5% improvement. Among the different classes, pedestrian detection proved to be the most challenging one with a 78% *F-score*.

We evaluated our 3D CNN-based semantic point cloud classification technique (Sec. 6.4.4) on real MLS data captured with a Riegl VMX-450 system in Budapest. The available measurement set contained in total around 300 Million points from various urban scenes, including main roads with both heavy and solid traffic, public squares, parks, and sidewalk regions, containing various types of cars, trams and buses, several pedestrians and diverse vegetation. As reference technique, we trained a single channel 3D CNN model [190], referred as OG-CNN, which used a 3D voxel occupancy grid as input feature. As metrics we calculated the voxel level *precision*, *recall* and *F-score* for each class separately as well as the overall performance. By analyzing the results, we could conclude that the proposed two-channel 3D CNN can classify all classes of interest with an *F-score* larger than 83%. The overall results of the reference OG-CNN technique fall behind our proposed method with 13%.

We validated the proposed multimodal point cloud registration process (Sec. 6.4.5) with matching the measurements of RMB Lidar sensors to the MLS point clouds. Since *Ground Truth*

transformation was not available, we calculated first an asymmetric Modified Hausdorff Distance (MHD) between the RMB Lidar and MLS clouds. We have observed that both the object based Hough matching, and NDT-based registration refinement steps [20] could significantly decrease the distances between the scans in almost all data sets. By estimating the accuracy of the technique, we had to eliminate the effects of several moving objects (especially large trams or tracks) which mislead the calculated MHD distances. Therefore, we also used a modified error metrics called Median Point Distance (MPD) [20], which estimated a median of point-level errors instead of averaging them. In seven out of the eight scenes the resulting MPD errors were below 3cm, which fact was also confirmed by visual verification (see details in Appendix E).

Since we have not found any similar RMB Lidar-MLS multimodal change detection approaches in the literature to our one presented in Sec. 6.4.6, we adopted a voxel based technique [191] as reference, which was originally constructed for already registered MLS/TLS point clouds. We used a voxel size of 0.3m for [191], which choice yielded approximately the same computational cost as our proposed MRF-range image based model (around 80msec/frame on a desktop computer, with CPU implementation). Comparative tests showed that the proposed method had an efficient overall performance, and it outperformed the voxel based method in general with 1-6% *F-score* in the different scenes. We have experienced that the main advantage of the proposed technique was the high accuracy of change detection in cluttered street regions, such as sidewalks with several nearby moving and static objects, where our method surpassed the voxel based approach with 7-15% gaps in the test scenes (see Table 6.5(b)).

6.5 Conclusions of the chapter

This chapter presented new methods for three different problems in 4D environment perception. *First*, we formulated 3D object detection in multi-camera systems as an inverse problem, and proposed a Marked Point Process based solution, which surpassed the state-of-the-art. *Second*, we have shown that the recently appeared Rotating Multi-beam (RMB) Lidar technology can also be utilized in advanced surveillance systems for people tracking, biometric re-identification and activity recognition. *Finally*, we have proposed a novel set of methodologies for real time Lidar-based urban scene analysis from a moving platform, relying on a reference 3D background map generated from mobile laser scanning data.

Chapter 7

Conclusions of the thesis

This thesis has focused on various region level and object based pattern recognition problems, which raise nowadays important challenges to experts in computer vision and machine perception. By most of the selected issues, stochastic Bayesian energy minimization techniques or supervised data classification approaches have been chosen as bases for the introduced new methods, and improvements versus state-of-the-art approaches have been proposed in various aspects, including observation processing, combination of different model structures, and new spatial and temporal interpretation of up to-date sensor measurements. While in several real time applications, the high computation cost of energy minimization methods may mean bottleneck of applying complex models, I have shown that with using appropriate dimension reducing techniques, combining stochastic and deterministic relaxation approaches, and the utilization of prior knowledge based rules we can often obtain high quality and computationally efficient solutions. Although the application examples from the thesis cover a broad field, I have mainly focused on general problems appearing concurrently in various domains, with exploring the possible applicability of the presented models under varying circumstances. The thesis put particular attention on the connections between the theoretical results of established mathematical models and the applicability of the implemented methods using real world data collected from realistic scenarios. For this reason, the validation of the proposed models has mostly been experimental, and significant efforts have been conducted for test data collection, Ground Truth generation and relevant quantitative comparison to state-of-the-art approaches targeting the same or similar problems.

7.1 Methods used in the research work

In the course of the my work, theorems and assertions from the fields of mathematical statistics, probability theory, optimization and reported results of image and video processing, point cloud analysis, data fusion and 3D/4D machine vision were explored.

The primary goal of the conducted research has been to extend the available libraries of computer vision and pattern recognition techniques with new complex but still general approaches, which can be widely applicable to solve real-life automatic perception problems. A special attention has focused on temporal and spatial multi-level decomposition of scenarios and events.

From a methodological point of view, several proposed methods are different implementations of various well established probabilistic models, such as *Markov Random Field*s (MRF, [102]), *Mixed Markov Models* [75], and *Marked Point Processes* (MPP, [83]). In this way I could fully take the advantages of a solid theoretical background of the models, which fact provided guarantees for stability, and facilitated estimating the domains of validity and limitations of the new approaches.

Since the main contributions of this dissertation consist of various modeling and algorithmic planning steps, thoughtful experimental evaluation was a critical issue to measure the significance of the improvements in the different application domains. I paid therefore particular attention to rely on relevant datasets and Ground Truth (GT) information either from publicly available databases, or from our own measured/obtained data samples. In the Machine Perception Research Laboratory of MTA SZTAKI various leading-edge sensors were available, including high resolution optical cameras, depth and thermal cameras and a Velodyne HDL 64-E terrestrial Lidar scanner, which can be mounted onto the top of a car. On the other hand, for remote sensing problems the test data was provided by our research partners: aerial & satellite images and Lidar scans from Budapest were obtained from Astrium Defense and Space Hungary Ltd, we received the radar (ISAR) image sequences from the University of Pisa, and mobile laser scanning (MLS) data from Budapest Közút Zrt. Some of the aerial photos were purchased from the Hungarian Institute of Geodesy, Cartography and Remote Sensing (FÖMI).

Most of our proposed new benchmarks were published in the website of our research team, thus the international scientific community might use them for comparative evaluation.

For the design and testing of algorithms I have mainly used C/C++software development environments, while some initial prototyping steps have been coded in Matlab. Implementing image processing and point cloud management/visualization routines in C++have been highly facilitated by the publicly available OpenCV [192] and PCL [188] software toolboxes.

7.2 New scientific results

The scientific contributions of the dissertation are presented in four thesis groups. In the *first thesis group*, we introduce new techniques for pixel wise change detection in remotely sensed images from two different application environments, using new multi-layer Markov Random Field models. The *second thesis group* focuses on object level change analysis with novel spatio-temporal marked point process models. In the *third thesis group* we extract (spatially) hierarchical object structures from digital images by a new embedded marked point process model. Finally in the *fourth thesis group* we provide solutions for different environment analysis tasks using novel sensor technologies, including person localization and biometric identification in videos surveillance systems, and dynamic environment perception from moving platforms with geographic information database background.

In the research works connected to Thesis groups 1-3, I worked as principal researcher being responsible for the definition of the exact research goals, model construction, literature review, and most of the implementation and testing issues. My co-authors mainly contributed to the model formulation and presentation parts with their expertize in probabilistic modeling, information fusion and image based change detection, and/or to problem definition with their application specific knowledge in remote sensing, radar imaging and industrial production technology. Results of Thesis group 4 were obtained in cooperation with postdoctoral colleagues and my supervised Ph.D. and undergraduate students, the share of contributions are detailed by the corresponding thesis points.

Thesis group 1: I have proposed new multi-layer Bayesian label fusion models for various change detection problems in remotely sensed images. I have introduced efficient optimizer algorithms for the developed models based on the Modified Metropolis Dynamics. I have experimentally validated the algorithms on real remote sensing applications, and showed their advantages versus earlier solutions from the literature.

> *1.1. I have introduced a new three-layer Markov Random Field (L^3MRF) model for detecting the regions of independent object motions in high resolution image pairs captured from moving aerial platforms. I have experimentally shown that the proposed method outperforms previous models which use purely linear image registration techniques or local parallax removal, and also demonstrated the advantages of the new technique versus alternative feature fusion approaches.*

Moving object detection in image streams of aerial vehicles needs to solve a frame differencing task with accurate camera motion compensation. Relying on a robust but coarse 2D frame matching algorithm, the main challenge of the L^3MRF model has been to eliminate the registration errors and parallax distortions from the extracted motion mask. Assuming local distortions with a bounded magnitude, I have shown that frame differencing and local block correlation provide

efficient complementary features to remove registration artifacts. Thereafter, a new three-layer Markov Random Field model has been introduced for the problem, where the segmentation of the outer layers are based on the two different features, while the central layer represents the final change mask without direct links to the observations. Intra-layer connections ensure smoothness of the segmentation within each layer, while inter-layer links provide semantically correct labeling in the central layer via pure label fusion constraints.

Evaluation has been conducted on three different data sets containing in aggregate 83 real aerial image pairs with manually generated pixel level Ground Truth information. A detailed quantitative comparison has shown the advantages of the new approach versus five state-of-the art solutions for the same practical problem. In addition, the significance of the proposed label fusion model has been demonstrated by a methodological study, where under the same image feature selection, the L^3MRF technique has been compared to alternative 'observation fusion' and 'decision fusion' schemas.

The L^3MRF method was published in IEEE TRANS. IMAGE PROCESSING in 2009 [12], partial results were presented earlier in [46, 54]. Although the model was partially included in my Ph.D. dissertation [53], additional contributions have been inserted in [12] regarding the quality and stability analysis of the proposed technique, and a detailed comparative evaluation has been given here firstly versus other approaches from the state-of-the-art.

1.2. I have proposed a new four-layer Conditional Mixed Markov model (CXM), as a combination of a mixed Markov model and a conditionally independent random field of signals, for detecting relevant changes in registered aerial image pairs taken with the time differences of several years and in different seasonal conditions.

Automatic comparison of optical aerial images with large time gaps is a highly challenging pattern recognition task, since due to varying illumination condition, seasonal vegetation changes, and different camera sensors, the extracted low level image features (such as color value or texture) may be significantly different even in 'unchanged' geographic regions. Label fusion through context based feature selection is a natural idea to approach the problem. Although context dependent class models can be hardly encoded in conventional Markov Random Fields (MRFs) defined on static graphs, the recently introduced Mixed Markov models [75], which admit data-dependent links between the processing nodes, enable configurable structures in feature integration.

I have proposed a novel multi-layer model structure, called the Conditional Mixed Markov model (CXM), which extends the multi-layer label fusion framework of Thesis 1.1 with dynamic feature based connections following the Mixed Markovian approach. Here the different layer-regions are considered or ignored by the label fusion rules upon local statistical estimation of the reliability of their corresponding features.

I have demonstrated the efficiency of the CXM framework on the (above defined) long-term change detection problem for aerial images. The proposed model implementation integrates global intensity statistics with local correlation and contrast features. A global energy optimization process ensures simultaneously optimal local feature selection and smooth, observation-consistent segmentation.

For evaluation three sets of real optical aerial images were used, which have been provided by the Hungarian Institute of Geodesy, Cartography and Remote Sensing (FÖMI) and Google Earth. The test sets contain 13 - also manually evaluated - photo pairs, covering in aggregate around 17 km^2 areas, with time differences of 5 to 23 years between the corresponding shots. Comparative results versus four state-of-the art methods (published in top journals of the field) confirmed the superiority of the proposed model, while performance variation with respect to perturbation of selected parameters has also been investigated.

The CXM method was published in IEEE TRANS. GEOSCIENCE AND REMOTE SENSING in 2009 [13], while partial results were presented at the ICPR 2008 conference [45]. In a survey paper published in *ISPRS Journal of Photogrammetry and Remote Sensing* [4] (2015) we compared the results to various newer state-of-the art solutions, and the advantages of the approach have been demonstrated again. This survey paper [4] served as the basis of a successful project proposal submitted for the call for Research Groups with Significant International Impact (KH-17 call) of the *National Research, Development and Innovation Fund (NKFIA)* in 2017.

Thesis group 2: I have introduced spatio-temporal Marked Point Process (MPP) models for object level time sequence analysis, by completing conventional MPPs with a new temporal dimension. Selected remote sensing applications on object-based change detection and moving target analysis in image sequences have been developed and thoroughly evaluated to demonstrate the advantages of proposed approach.

Marked Point Processes have already been proved to be efficient tools in various object population counting problems dealing with a large number of objects which have low varieties in shape, such as buildings or trees from remotely sensed data, cell nuclei from medical images, or people in video surveillance scenarios. While previous attempts deal with static scenarios, many applications may require handling object level change detection or dynamic target surveillance problems in the advantageous geometric approach.

2.1. I have introduced a new probabilistic approach, called the Multitemporal Marked Point Process (mMPP) model, which integrates geometric object extraction with low level change detection in a joint framework. I have implemented and evaluated the model for the application of building change monitoring in aerial or satellite images. The advantages of the approach have been demonstrated over existing state-of-the-art techniques.

Following the evolution of built-up regions is a key issue of aerial and satellite image analysis. Focusing on this crucial applicational need, I have developed a new probabilistic method, which

presented methodological contributions in three key issues:

(i) I implemented a novel object-change modeling approach based on Multitemporal Marked Point Processes (mMPP), which simultaneously exploits low level change information between the time layers and object level building description to recognize and separate changed and unaltered buildings.

(ii) Answering the challenges of *data heterogeneity* in aerial and satellite image repositories, I constructed a flexible hierarchical framework which can create various building appearance models from different elementary feature based modules.

(iii) To simultaneously ensure the convergence, optimality and computation complexity constraints raised by the increased *data quantity*, I adopted the quick *Multiple Birth and Death* optimization technique for the change detection task, and proposed a novel non-uniform stochastic object birth process, which generates relevant objects with higher probability based on low-level image features.

The proposed model has been validated using eight significantly different aerial and satellite image data sets containing 662 manually labeled buildings. Object and pixel level quantitative evaluation results have been provided, and it has been shown that the building localization performance of the new approach outperforms four state-of-the-art techniques from the literature, with competitive figures regarding the computational time. From a methodological point of view, the efficiency of the proposed joint object-change classification framework has been demonstrated versus the conventional post detection comparison schema, while the advantages of the introduced feature-based birth process and the effects of various parameter settings have been deeply analyzed.

The publication introducing the mMPP model appeared as the featured article of the January 2012 issue of IEEE TRANS. PATTERN ANALYSIS AND MACHINE INTELLIGENCE [10], while corresponding results were presented in various conferences [40, 42, 43, 44].

2.2. I have introduced a dynamic Multiframe Marked Point Process (F^mMPP) framework for moving target analysis, which can be used for the simultaneous extraction and tracking of objects and characteristic feature points in noisy image sequences. I have demonstrated the efficiency of the approach for automatic target structure extraction and tracking in the series of Inverse Synthetic Aperture Radar (ISAR) images, where due to focusing artifacts of image formation and strong speckle noise effects object fitting on independent frames is largely unreliable. I have shown that using the F^mMPP approach we can obtain a robust result for the whole sequence by exploiting prior geometric constrains between the neighboring frames regarding target position estimation and shape consistency.

Identification and motion analysis of ship targets in airborne Inverse Synthetic Aperture Radar (ISAR) image sequences are key problems of Automatic Target Recognition (ATR) systems which

utilize ISAR data. Remotely sensed ISAR images are able to provide valuable information for target classification and recognition in several difficult situations, where more traditional SAR imaging techniques fail. However, robust feature extraction and feature tracking in ISAR image sequences are usually difficult tasks due to noise and the low level of available details about the structure of the imaged targets.

I have proposed a new Multiframe Marked Point Process (F^mMPP) model of line segments and point groups for automatic ship and airplane structure extraction and target tracking in ISAR image sequences. A robust joint model has been developed for axis extraction, feature point detection and tracking. For the purpose of dealing with scatterer scintillations and high speckle noise in the ISAR frames, the resulting target sequence has been obtained by an iterative optimization process, which simultaneously considered the observed image data and various prior geometric interaction constraints between the target appearances in the consecutive frames.

Quantitative evaluation has been performed on 8 real ISAR image sequences of different carrier ship and airplane targets, using a test database containing 545 manually annotated frames. I have experimentally shown that in case of noisy sequences, the introduced F^mMPP schema can significantly improve the results of the frame-by-frame detection steps.

The proposed approach was published in IEEE TRANS. GEOSCIENCE AND REMOTE SENSING in 2014 [7], and partially presented in a radar-focused [33] and in a general remote sensing conference [38].

Thesis group 3: I have introduced the three-layer Embedded Marked Point Process (EMPP) framework for extracting complex hierarchical object structures from various digital images used by machine vision applications. The proposed method has been demonstrated in three different application areas: optical circuit inspection, built in area analysis in remotely sensed images, and traffic monitoring on airborne Lidar data.

Classical MPP-based image analysis models focus purely on the object level of the scene, and they cannot be suited to hierarchical pattern recognition problems in a straightforward way.

For overcoming the above limitations, I have proposed the EMPP framework, which extends conventional Marked Point Process models by admitting object-subobject ensembles and allowing corresponding objects to form coherent object groups. These two contributions were initially motivated by definite practical requirements from optical circuit inspection (Thesis 3.1), and remote sensing traffic monitoring [5] applications. After solving tasks specific problems, I have defined and implemented a general three layer model framework (Thesis 3.2), which has been simultaneously tested and validated in three significantly different domains.

3.1. I have introduced an automated Bayesian visual inspection method for Printed Circuit Board (PCB) assemblies, which is able to simultaneously deal with variously shaped Circuit Elements (CE) on multiple scales, by including object-subobject ensembles in the Marked Point Process schema. I have demonstrated the efficiency of the approach on the task of

solder paste scooping detection and scoop area estimation, which are important factors in PCB quality inspection.

Automatic optical inspection (AOI) technologies provide very high resolution (10μm) images of printed circuit boards (PCB), thus a comprehensive quality analysis needs a hierarchical modeling approach of the PCB structure, focusing jointly on circuit regions, individual Circuit Elements (CEs), CE interactions and searching for characteristic patterns within the CEs, like the geometric scooping artifacts.

I have proposed a new visual inspection method, which describes the hierarchy between objects and object parts as a parent-child relationships embedded into the MPP framework. To simultaneously deal with variously shaped circuit elements, different types of geometric objects are jointly sampled, by adopting the multi-marked point process schema to the hierarchical entity extraction problem. Since due to these methodological modifications, the dimension and size of the solution space has been significantly increased compared to conventional MPP models, it became crucial to efficiently sample the population space during model optimization. Therefore, by extending the non-uniform object birth process from Thesis 2.1, I have developed a *Bottom-Up (BU) stochastic object proposal strategy*, by combining low level statistical image descriptors with prior information based structure estimation, which step has kept the computational complexity tractable.

The proposed method has been evaluated on real PCB data sets containing 125 images with more than 10.000 circuit elements. Performance efficiency has been demonstrated versus a conventional morphology based fault detection technique.

The core of the proposed pattern recognition approach was published in a sole author paper in *Pattern Recognition Letters* [11] and presented at the IEEE ICIP 2011 conference [37], while the complete model was introduced in IEEE TRANS. INDUSTRIAL ELECTRONICS in 2013 [8], where the technological background of the selected AOI problem was provided by expert co-authors of electronic technology from the Budapest University of Technology and Economics.

3.2. I have defined a general three-layer Embedded Marked Point Process (EMPP) model with a corresponding multi-layer layer energy optimization algorithm, which can simultaneously extract object groups, objects and object parts from high resolution digital images. With ensuring flexible designing options of the data based and prior constraints in the model, I have shown that the EMPP approach can be fit to various real world hierarchical pattern recognition problems. The performance of the new technique has been validated in three different application domains.

A number of techniques have been previously proposed for multi-level content analysis of high resolution images, following either region based [71], (the above introduced) object based [5, 8], or hybrid [72] approaches. However, the these models were suited to specific application areas with specific inputs: remotely sensed optical images [71, 72], Lidar point clouds [5], and - as shown in Thesis 3.1 - PCB AOI tasks using μm resolution images [8]. Experiences show that for

such complex, application dependent models, the adaption to another application domain is rarely straightforward, needing a significant modeling and implementation work.

For this reason, taking a reverse approach, I collected similar problems from my previously analyzed application fields, and proposed a novel general three-level Embedded Marked Point Process (EMPP) framework which can handle a wide family of applications. The structure elements and the energy optimization algorithm of the complex model were defined and implemented at the abstract level, while I kept focus on ensuring very simple interfaces to the different applications, enabling efficient domain adaption.

The key contributions of the proposed EMPP methodology over the conventional single layer MPP models are as follows:

(i) We describe the hierarchy between objects and object parts as a parent-child relationship embedded into the MPP framework. The model of a child is affected by its parent entity, considering geometrical and spectral constraints.

(ii) We partition the (parent) entity population population into object groups, called configuration segments, and extract the objects and the optimal segments simultaneously by a joint energy minimization process. We create adaptive object neighborhoods by segment driven object interactions.

The proposed method has been demonstrated in three different application areas: built in area analysis in remotely sensed images, traffic monitoring on airborne Lidar data and optical circuit inspection. In addition, a detailed methodological validation process has been conducted: I have quantitatively demonstrated the advantages of the EMPP approach versus a sequential technique, which extracts first the object population by a single layer MPP model (sMPP), thereafter, the parent object grouping step is performed in post processing by a recursive floodfill-like segmentation of the population. I have also performed a detailed analysis on the repeatability of obtained results using the iterative optimization process, and on the computational time requirements of the algorithm.

The general EMPP model has been published in a sole author paper in IEEE TRANS. IMAGE PROCESSING in 2017 [2], and presented at ICASSP 2014 [25] and ICIAR 2013 [28] conferences.

Previously, a two-level, applications specific version (called L^2MPP) focusing on the extraction of vehicles and coherent vehicle groups from aerial Lidar data has been published in IEEE TRANS. GEOSCIENCE AND REMOTE SENSING in 2015 [5], and some preliminary model elements have been presented at the ICPR 2012 [32] and ISPRS 2012 [34] conferences. The proposal of the L^2MPP model also contained major contributions of my Ph.D. student Attila Börcs [193], especially in the point cloud segmentation and in the feature-based energy model construction

parts, while I have significantly contributed here to the construction and evaluation of the prior data model and the optimization algorithm.

Thesis group 4: I have proposed new models and algorithms contributing to various video surveillance and environment perception tasks based on 4D spatio-temporal measurements of heterogeneous sensors. The algorithms have been quantitatively evaluated on representative real world data sets, and the performance improvement versus existing state-of-the-art solutions has been demonstrated.

Object localization, classification, motion tracking and change detection are important issues in intelligent surveillance and environment perception applications, such as security surveillance, autonomous driving, or city management. However, these tasks are still challenging in crowded outdoor scenes due to uncontrolled illumination conditions, irrelevant background motion, and high occlusion rates caused by several moving and static scene objects. Involving novel 3D/4D sensory information may significantly contribute to the workflow, however it also implies new challenges for machine vision technologies. I investigated selected issues in this problem domain, relying on various hardware configurations.

4.1. I have proposed a Marked Point Process Model of cylinders for modeling groups of multiple (possibly overlapping) pedestrians in 3D environments. Using features extracted from different calibrated camera views of a multi-camera system, it has been shown that the introduced approach can be efficiently applied for accurate 3D localization and height estimation of people, surpassing a state-of-the-art solution for the same problem.

In this work a Bayesian approach has been proposed on multiple people localization in multi-camera systems. First, pixel-level features have been extracted by my co-author, which were based on physical properties of the 2D image formation process, and provided information about the head and leg positions of the pedestrians, distinguishing standing and walking people, respectively. Then features from the multiple camera views have been fused to create evidence for the location and height of people in the ground plane. This evidence accurately estimated the leg position even if either the area of interest were only a part of the scene, or the overlap ratio of the silhouettes from irrelevant outside motions with the monitored area were significant.

Using the above feature information, I have defined a 3D object configuration model in the Euclidean coordinate system of the scene. I also utilized prior geometrical constraints, which describe possible interactions between pairs of neighboring pedestrians. To approximate the position of the people, I used a population of 3D cylinder objects, which was realized by a Marked Point Process. The final configuration results were obtained by a conventional iterative stochastic energy optimization algorithm.

The proposed approach has been evaluated on two publicly available datasets, and compared to a recent state-of-the-art technique. To obtain relevant quantitative test results, 3D Ground Truth

annotation of the real pedestrian locations has been prepared, while two different error metrics and various parameter settings were evaluated, showing the advantages of our proposed model.

The proposed 3D MPP model has been published in IEEE TRANS. CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY in 2013 [9], presented at the CVPR 2011 [39] premier computer vision conference (acceptance rate was 26%), and in specific workshops of the topic [35, 36, 41].

4.2. I have proposed novel algorithms for people surveillance -including motion detection, tracking and event analysis - using a single rotating multi-beam Lidar sensor, which monitors a dynamic scene from a fixed position. The proposed methods have been quantitatively evaluated in realistic surveillance and traffic monitoring environments, showing that the proposed approaches outperform concurrent state-of-the-art methods.

While conventional optical or range sensors have a limited Field of View (FoV), Rotating Multi-beam (RMB) Lidars – used as surveillance cameras – provide a full 360° FoV of the scene, with a vertical resolution equal to the number of the sensors, while the horizontal angle resolution depends on the speed of rotation. However, inhomogeneous point cloud density, noise of sensor calibration and consequences of the high speed sequential scanning process introduce various artifacts for the data processing modules, demanding novel solutions in data filtering and pattern recognition. My key developments in this environment were as follows:

(i) I introduced a hybrid 2D–3D approach for dense foreground-background segmentation of RMB Lidar point cloud sequences obtained from a fixed sensor position. The proposed technique solved the computationally critical spatial filtering steps in the 2D range image domain by a Markov Random Field (MRF) model, however, ambiguities of discretization were handled by joint consideration of true 3D positions and backprojection of 2D labels. By adopting a spatial foreground model (originally introduced in my Ph.D. dissertation [53]) to the range image domain, I could significantly decrease the spurious effects of irrelevant background motion, which was principally caused by moving tree crowns and bushes. For quantitative point level evaluation of the Lidar scenario, a 3D point cloud Ground Truth (GT) annotation tool has been developed. The proposed foreground extraction module has been compared to various alternative state-of-the-art techniques and its superiority has been demonstrated both in people surveillance and in terrestrial traffic monitoring scenarios.

(ii) I proposed a real-time method for moving pedestrian detection and tracking in RMB Lidar sequences for dense surveillance scenarios, with short- and long-term object assignment. During the Short-Term Assignment (STA) the different people were separated in the foreground regions of the point cloud frames, and the corresponding centroid positions were assigned to each other over the consecutive time frames. The Long-Term Assignment (LTA) was responsible for connecting the broken trajectories caused by STA errors and identifying the re-appearing people.

(iii) I introduced the utilization of gait based biometrics for the RMB Lidar sequences to support the LTA step of the above tracker. Various alternative features have been extracted, tested and quantitatively compared in cooperation with my co-authors, yielding the superiority of a new Lidar based Gait Energy Image (LGEI) descriptor. We have also proposed the extension of the LGEI approach to recognize various typical activities (such as bending or waving), which has been successfully implemented and evaluated during the research work.

The complete workflow of Lidar-based surveillance framework with a main focus on person re-identification and activity recognition was published in IEEE TRANS. CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY in 2018 [1]. The 3D people surveillance approach was first introduced in my sole author paper in *Pattern Recognition Letters* in 2014 [6], The foreground-background segmentation algorithm has been initially presented at the WDIA 2012 ICPR Workshop [31], while preliminary results in gait recognition were demonstrated at the EUSIPCO 2015 [22], IWCIM 2015 [21] and VISAPP 2017 [19] conferences.

4.3. I have proposed a novel workflow for the analysis of dynamic urban environment using RMB Lidar point cloud sequences captured from a moving vehicle, and very dense reference background point clouds obtained with mobile laser scanning (MLS) technology. I have contributed to the development and validation of novel algorithms for object detection, classification, background scene segmentation, multimodal point cloud registration and change detection in the proposed framework.

Laser scanning technologies can provide high precision 3D measurements form the environments. Using up-to-date sensors, however, we still find a trade-of between the spatial and temporal resolution of the recorded point cloud data. Rotating Multi-beam (RMB) Lidar sensors mounted on vehicle tops can collect point cloud sequences with 15-30 frames per second, allowing dynamic event analysis, however as mentioned by Thesis 4.2, the spatial density of the measurements is low and strongly inhomogeneous. On the other hand, mobile laser scanning (MLS) systems can produce very high resolution static 3D point clouds of large scale city regions, which contain apart form the really *stationary* scene parts (e.g. road, facades, lamp posts) various *movable* (e.g. parking cars) and *changing* (vegetation) objects, and *phantom* regions caused by objects concurrently moving with the scanning platforms.

I have proposed a workflow of algorithms for facilitating the joint exploitation of the measurements from RMB Lidars in the cars' instant sensing platforms and offline spatial databases containing geo-referred MLS point cloud data. With a team of my undergraduate and Ph.D. students, we have been working towards and a new algorithmic toolkit which allows self driving cars to obtain in real time relevant Geographic Information System (GIS) information for decision support, and provides opportunities for extending and updating the GIS databases based on the sensor

measurements of the vehicles in the everyday traffic. We have proposed a set of new algorithms detailed as follows.

We have introduced a real time method for ground-obstacle segmentation of RMB Lidar point clouds with the separation of *low foreground* (regions of short street objects) and *high foreground* (building facades and tall objects) areas, considering the inhomogeneous spatial density of the RMB Lidar frames, uneven road surfaces and the presence of dense traffic in urban environments [30]. Based on the above segmentation, we have proposed a two-level grid based quick object extraction algorithm working in the low and high foreground regions, which enables the real time separation of several nearby objects [24], and efficient 2D top-view bounding box fitting [23] using structural constraints. Next, the low foreground objects have been classified into four semantic classes: vehicle, pedestrian, street clutter and short facade, in a range image based representation, using a (2D image based) Convolutional Neural Network (CNN) classifier [3].

We have also proposed a 3D voxel based CNN approach for semantic segmentation of the MLS data [18], which can be used to remove *phantom* regions and movable objects from the point clouds, and mark the static landmark objects, which can be used by self driving vehicles for orientation in the 3D high definition map. Next we introduced a multimodal pont cloud registration algorithm, which enables the accurate positioning of actual RMB Lidar point cloud frames in the MLS map's global coordinate system [16, 20]. Finally, we worked out a Markov Random Field based change detection technique between the registered multimodal point clouds [17].

The RMB Lidar based object detection and classification approach has been published in IEEE GEOSCIENCE AND REMOTE SENSING LETTERS [3] and in a Springer book chapter [50], while the preliminary and later results of this topic have been presented in various conferences: ISPRS VMC 2013 [30], Workshops of ECCV 2014 [24], ACCV 2014 [23] and ECCV 2018 [16], at ICPR 2016 [20], IJCNN 2017 [18] and ICIAR 2017 [17].

In the work presented in this sub-thesis, I participated as principal investigator and coordinator, specifying the workflow and the major steps of the research work, and supervising the research and development steps. Most of the specific technical contributions are shared with *Ph.D.* and undergraduate students working under my supervision, in particularly with Attila Börcs (Ph.D., 2018, [193]), Balázs Nagy (Ph.D. studies in progress), Bence Gálai (M.Sc., 2017) and Oszkár Józsa (B.Sc., 2013).

7.3 Examples for application

The developed algorithms can be used by various up-to-date or future computer vision systems, especially in the application fields of video surveillance, remote sensing, industrial quality analysis, film pre-production, robotics and autonomous driving etc. Many of the proposed methods directly

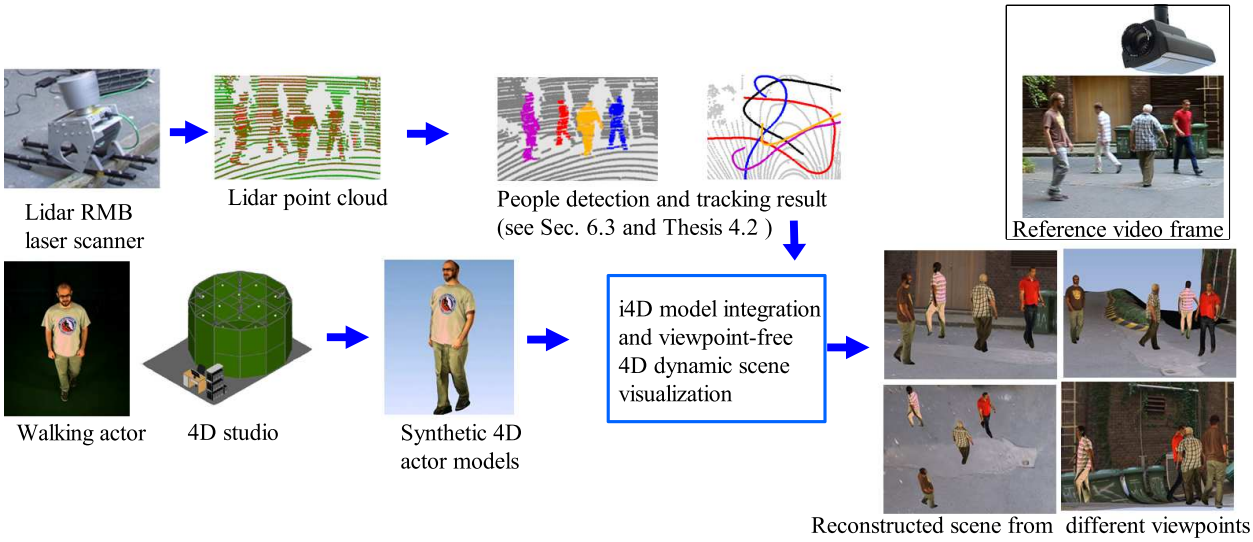


Figure 7.1: Flowchart of the i4D system, patented in [48]

corresponded to research projects conducted with the participation of MTA SZTAKI in the previous years. My scientific contributions were in particularly involved in the following projects, where I also acted as a coordinator, principal investigator or (national) project leader.

(i) The Array Passive ISAR Adaptive Processing (APIS) project of the European Defense Agency (EDA) focused on the development and demonstration of the functionalities of a passive high-resolution primary radar system. The result of APIS was to our best knowledge the first passive system in the world that was capable of offering images with the application of Inverse Synthetic Aperture Radar (ISAR) techniques. I was working in the project as the local technical coordinator from the side of MTA SZTAKI and a participating senior researcher, responsible for image analysis and pattern recognition tasks of the project. My contributions regarding moving target analysis in ISAR image sequences [7] (introduced in Thesis 2.2), were involved in significant parts in the final project report and in the demonstrator.

(ii) I have coordinated the Integrated 4D (i4D) research project at MTA SZTAKI, which developed an unconventional hardware-software environment, combining two very different sources of spatiotemporal information: a RMB Lidar and a 4D reconstruction studio (see Fig. 7.1). The main scientific purpose of the integration of the two types of data has been the desire to measure and represent the visual world at different levels of detail. In the proposed approach, the Lidar sensor provided a global description of a dynamic outdoor scene in the form of a time-varying 3D point cloud. The latter was used to separate moving objects from static environment and obtain a 3D model of the environment. The 4D studio built a detailed dynamic model of an actor (typically, a person) moving in the studio. By integrating the two

sources of data, which was to our best knowledge a unique attempt up to now, one could modify the model of the scene and populate it with the avatars created in the studio. We have addressed various application areas such as 4D virtual city reconstruction, protecting collective properties in urban environments, 4D video surveillance, augmented reality, and telecommunication. We described the main scientific and technical novelties of the i4D system in a patent [48], and published at referred conferences IEEE Coginfocom 2013 [26], IEEE CBMI 2013 [29] and ICVS (LNCS) 2013 [27]. Based on the i4D technology, we developed a prototype system supporting 4D film preproduction, which was demonstrated at the FMX 2017 Visual Effect Conference and Exhibition, and was selected to appear in the book *100 most interesting Hungarian inventions* in 2018.

(iii) I was the coordinator and a research scientist in the DUSIREF (Dynamic Urban Scene Interpretation and REconstruction through remotely sensed data Fusion) Project funded by the European Space Agency (ESA) under the PECS-HU framework. The main objective of the project was high level urban scene recognition and change interpretation based on heterogeneous Remote Sensing (RS) data sources (mainly optical and TerraSAR satellite images and LIDAR data). We aimed to develop novel recognition and visualization methodologies relying on four dimensional data representation, focusing on highly multi-modal, multi-scaled and multi-temporal data collections: here my contributions introduced in Thesis 1.2, 2.1 and 3.3 were widely exploited.

(iv) I am currently the principal investigator of two exploratory research projects funded by the National Research, Development and Innovation Fund (NKFIA), with titles *Instant environment perception from a mobile platform with a new generation geospatial database background* (K-16 call for Researcher Initiated Projects: 2016-2020), and *Change detection and event recognition with fusion of images and Lidar measurements* (KH-17 call for Research Groups with Significant International Impact, 2017-2019), whose scopes largely overlap with the research work presented in Theses 4.2 and 4.3.

My further scientific results were also utilized by major or minor components of various project contributions delivered by MTA SZTAKI, Péter Pázmány Catholic University (PPCU) or Budapest University of Technology and Economics (BME). The video surveillance methods from Thesis 4.1 were used by EDA Project MEDUSA and EU Project THIS. The proposed circuit inspection technology (Thesis 3.1) was connected to the scientific program of the "Development of quality-oriented and harmonized R+D+I strategy and functional model at BME" project. The object motion detection technique for aerial image sequences (Thesis 1.1) was utilized by the *Hungarian R&D Project ALFA* (NKFP 2/046 /04 project funded by NKTH). Our geospatial data processing

Figure 7.2: Live demonstration of our Lidar-based person tracker at the Frankfurt Motorshow 2017, in the exhibition area of Velodyne

algorithms are applied for archeological data analysis in the EFOP–3.6.2–16–2017–00013 project at PPCU. Various contributions in RMB Lidar based scene analysis have been adopted in automotive industrial projects.

We also integrated the person surveillance module of Thesis 4.2 into a real time demonstrator, which has been introduced at the Frankfurt Motorshow 2017 (see Fig. 7.2) in the exhibition area of sensor producer *Velodyne*, at the Automotive Hungary 2017 exhibition, and in multiple Researchers' Night occasions (in Hungarian: *Kutatók éjszakája*), which are open yearly events for the public to visit research centers in Hungary.

7.4 Lecturing and domestic publications

Besides my research activities conducted in MTA SZTAKI, I was appointed as *associate professor* at PPCU in 2015, where I am responsible for the courses Computer Graphics and Basic Image Processing. I defended my habilitation thesis [49] in 2017. My supervised undergraduate and doctoral students from PPCU and BME won several awards, including *four* first prizes at national scientific student conferences (OTDK), Attila Kuba Prize, Dénes Gábor scholarship, national B.Sc. thesis competition and various SZTAKI awards.

I am a steering committee member of the Hungarian Association for Image Analysis and Pattern Recognititon (*KÉPAF*) since 2013, and the *president* of the society since January 2019. Several publications connected to the dissertation appeared also in Hungarian language in the Proceedings of the *KÉPAF* Conferences, including our contributions on multi-layer Markov models [70], building change detection [69], radar image sequence analysis [65], Embedded MPPs [60, 66], multi-camera person localization [68], Lidar-based surveillance [57, 62, 67], and 4D environment analysis [55, 56, 58, 59, 61]. I also presented papers in the Hungarian Computer Graphics and Geometry Conference (GRAFGEO) [63, 64].

Appendix A

Summary of abbreviations and notations

List of abbreviations and concepts	
Abbreviation	**Concept**
MRF	Markov Random Field
CXM	Conditional Mixed Markov Model
DMRF	Dynamic Markov Random Field
MPP	Marked Point Process
mMPP	Multitemporal Marked Point Process (used for change detection)
F^mMPP	Multiframe Marked Point Process (object sequence analysis)
EMPP	Embedded Marked Point Process (multi-level MPP)
MAP	Maximum a posteriori
SA	Simulated Annealing (optimization method)
pdf	probability density function
MMD	Modified Metropolis Dynamic (MRF SA relaxation technique)
MBD	Multiple Birth and Death (MPP SA optimization technique)
RJMCMC	Reversible Jump Markov Chain Monte Carlo (MPP optimization)
PCA	Principal Component Analysis
PCC/PDC	Post Classification/Detection Comparison
RANSAC	Random sample consensus
FBB	Feature Based Birth Process
BUSEP	Bottom-Up Stochastic Entity Proposal
M^MBDM	Multi-level Multiple Birth-Death-Maintenance (MBD extension)
GODH	Gradient Orientation Density Histogram
ISAR	Inverse Synthetic Aperture Radar
PCB	Printed Circuit Board
AOI	Automated Optical Inspection

List of abbreviations and concepts	
Abbreviation	**Concept**
POM	Probabilistic Occupancy Map (multiview technique)
MLS/TLS	mobile/terrestrial laser scanning
Lidar	Light detection and ranging
RMB Lidar	Rotating Multi-beam Lidar (sensor)
ToF	Time-of-Flight
HD map	High Definition map
MLP	Multi-Layer Perceptron
CNN	Convolutional Neural Network
MoG	Mixture of Gaussians
MDA	Multiple Discriminant Analysis
(L)GEI	(Lidar-based) Gait Energy Image
STA/LTA	Short/Long Term Assignment
FoV	Field of View
ICP	Iterative Closest Point algorithm
NDT	Normal Distribution Transform algorithm
MHD	Modified Hausdorff Distance
e.g.	for example (in *latin:* 'exempli gratia')
i.e.	That is; in other words

General notations used in the thesis	
Variable	**Definition**
i, j, k, m	arbitrary index (number or enumeration)
n	dimension parameter, index
S	pixel lattice
s, r	pixel of the image lattice ($s, r \in S$)
$\mathcal{L} = \{p_1, \ldots, p_n\}$	point cloud of n points (p_i is the ith point)
x, y, z	Cartesian point coordinates
$G(\leftarrow S)$, G_i	image (over S lattice), the ith image
$g(s)$, $g_i(s)$	gray value/image sensor value at pixel s (in the ith image)
$N(\mu, \sigma)$	normal distribution with mean value μ and standard deviation σ
$\eta(x, \mu, \sigma)$	Gaussian (normal) pdf with parameters μ, σ.
$B(x, \alpha, \beta)$	beta pdf with parameters α, β.
$\zeta(x, \tau, m)$	sigmoid function with parameters τ and m.
κ_s, κ_s^i	weight (of the ith term) in a mixture pdf corresponding to pixel s
t, $\cdot^t$, $_t$	time (upper or lower) index (for any quantities)
T	temperature (for simulated annealing)
$\mathfrak{T}$	transform

Specific notations used in MRF/CXM models	
Variable	**Definition**
$\mathcal{G}(\mathcal{V}, E)$	MRF graph with set of nodes $\mathcal{V}$ and edges E.
v	abstract node of a graph $\mathcal{G}$, $v \in \mathcal{V}$ (without emphasizing which is the corresponding pixel in the input image)

Specific notations used in MRF/CXM models	
Variable	**Definition**
ε	edge of a graph $\varepsilon \in E$
s, r	node of $\mathcal{G}$ in case of a single layer MRF model
s^i	node at the i layer, which corresponds to pixel $s \in S$ (in case of multi-layer MRF models)
Λ	label set ($\#\Lambda = J$)
l, l_i	abstract label or class identifier
$\mathcal{N}$	neighborhood system of $\mathcal{G}$
$\mathcal{N}_v$	neighborhood of node v in $\mathcal{G}$ ($\mathcal{N}_v \in \mathcal{N}$)
$\varsigma(v)$, ς_v	label of node v in $\mathcal{G}$ ($\varsigma(v) \in \Lambda$).
ϖ	global labeling: $\{[v, \varsigma(v)] \vert v \in \mathcal{V}\}$
$\widehat{\varpi}$	MAP estimation of the optimal global labeling
$\varsigma^\star(a)$	the label of the regular node addressed by a in CXM
Υ	set of all the possible global labelings ($\varpi \in \Upsilon$)
ϖ_X	label subconfiguration corresponding to set $X \subseteq \mathcal{V}$ ($\varpi_X \subseteq \varpi$)
$V_X(\varpi)$	potential of the subconfiguration ϖ_X
$f(s)$, $\overline{f}(s)$	observation vector ($\in \mathbb{R}^n$) at pixel $s \in S$
$f(v)$, $\overline{f}(v)$	observation vector ($\in \mathbb{R}^n$) assigned to node $v \in \mathcal{V}$
$f_i(v)$ $\big(f_i(s)\big)$	ith component of vector $f(v)$ ($f(s)$)
$\mathcal{F}$	global observation on $\mathcal{G}$: $\{f(v) \mid v \in \mathcal{V}\}$
C	clique of $\mathcal{G}$
$\mathcal{C}$	set of cliques in $\mathcal{G}$
V_C	potential of clique C
$V_{\{v_1,\ldots,v_n\}}$	potential of a clique containing nodes $v_1, \ldots, v_n$
$\Theta(\varsigma_1, \varsigma_2)$	Potts smoothing term
δ, δ^i	parameter of the smoothing term (in the ith layer)
ρ	parameter of the inter-layer potential term
$\mathbf{1}\{E\}$, $\mathbf{1}_\varsigma$	indicator function of an event E, or class ς
$p_\varsigma(s)$	pdf value corresponding to pixel s and class ς
$\epsilon_\varsigma(s)$	$-\log p_\varsigma(s)$

Specific notations used in MPP models	
Variable	**Definition**
u, v	MPP objects represented by geometric figures
$\mathcal{H}$	parameter space of the objects
$\omega = \{u_1, \ldots, u_n\}$	configuration (or population) of n objects
Ω	configuration space
$\Phi(\omega) : \Omega \to \mathbb{R}$	MPP configuration energy function
$f(u)$ ($f^i(u)$)	data-feature associated to object u (i-named data-feature)
$\mathcal{F}$	global observation data over the input image(es)/point cloud(s)
$\mathcal{M}(f, d_0, D)$	feature-mapping function (d_0: acceptance threshold, D: normalization)

Specific notations used in MPP models	
Variable	**Definition**
$R_u \subset S$	set of image pixels covered by the geometric figure of object u
$\varphi_f(u)$	data-term of object u considering feature f
$A(u)$	unary potential of object u in $\Phi(\omega)$
$I(u, v)$	interaction potential between (parent) objects u and v
ψ	object group (in the EMPP model)
q (q_u)	child object (of parent u) in the EMPP model
Q_u	set of child objects of parent u (in the EMPP model)
$u \sim v$	neighborhood relation
$\mathcal{N}_u(\omega)$	neighborhood of object u within population ω: $\{v \in \omega \vert u \sim v\}$

Appendix B

Supplement regarding multi-layer label fusion models

This appendix offers supplementary materials regarding two issues of the multi-layer L^3MRF model, introduced in Sec. 3.2 for object motion detection in aerial image pairs.

First, the pseudo code of the proposed three-layer Modified Metropolis optimization technique is provided by Algorithm B.1.

Second, for qualitative comparison of the proposed L^3MRF model and the discussed reference techniques, Fig. B.1 shows four selected image pairs from the test database, segmented images with the different methods and Ground Truth change masks. Based on this figure and the quantitative test results (see Fig. 3.5 in Sec. 3.2.4), we can conclude that both the unsupervised *Reddy* [118] and the supervised *Affine* [123] methods cause many false positive foreground pixels due to the lack of parallax removal. The *Farin* model [121] can eliminate most of the misregistration errors got by [118], however, it may leave false foreground regions in areas with densely distributed edges and makes some small and low contrasted objects disappear. Since the Epipolar filter is based on local pixel correspondences, its artifacts may appear due to the failures of the feature tracker as well as in the case of objects moving in the epipolar direction [124]. During the tests of the *Epipolar method*, we have observed therefore both false alarms and missing objects (Fig. B.1). The bottleneck of using KNNBF [126] proved to be the poor quality of the region and application maps which could be extracted from the test images.

Algorithm B.1: Modified Metropolis algorithm used for the L^{3}MRF model

1. Pick up randomly an initial configuration ϖ, with $k := 0$ and $T := T_0$.
2. Denote by $|\mathcal{V}|$ the number of nodes in the three-layer model. Assign to each node a unique ordinal number between 1 and $|\mathcal{V}|$, applying the 'checkerboard' scanning strategy [107] for the consecutive layers. Let $j := 1$.
3. Let υ the j^{th} node, $i \in \{d, c, *\}$ is the layer which contains υ, while $s \in S$ is the corresponding pixel in the image lattice: $\upsilon = s^i$.
4. Denote the label of υ in ϖ by $\varsigma(\upsilon)$. Flip the label of υ and denote it by $\breve{\varsigma}(\upsilon)$.
5. Compute ΔU as follows:

$$\Delta U := \Delta U_1 + \Delta U_2 + \Delta U_3, \quad \text{where}$$

a. Calculate ΔU_1 as:

$$\Delta U_1 := \begin{cases} \log P\left(f_d(s)|\varsigma(\upsilon)\right) - \log P\left(f_d(s)|\breve{\varsigma}(\upsilon)\right) & \text{if } i = d, \\ \log P\left(f_c(s)|\varsigma(\upsilon)\right) - \log P\left(f_c(s)|\breve{\varsigma}(\upsilon)\right) & \text{if } i = c, \\ 0 & \text{if } i = *. \end{cases}$$

b. Using eq. (3.2), calculate ΔU_2 as:

$$\Delta U_2 := \sum_{r \in \mathcal{N}_s} \Theta\left(\breve{\varsigma}(s^i), \varsigma(r^i)\right) - \Theta\left(\varsigma(s^i), \varsigma(r^i)\right).$$

c. Denote by $V^0_{C_3} = V_{C_3}\left(\varsigma(s^d), \varsigma(s^c), \varsigma(s^*)\right)$ (eq. (3.4)). Calculate ΔU_3 as:

$$\Delta U_3 := \begin{cases} V_{C_3}\left(\breve{\varsigma}(\upsilon), \varsigma(s^c), \varsigma(s^*)\right) - V^0_{C_3} & \text{if } i = d, \\ V_{C_3}\left(\varsigma(s^d), \breve{\varsigma}(\upsilon), \varsigma(s^*)\right) - V^0_{C_3} & \text{if } i = c, \\ V_{C_3}\left(\varsigma(s^d), \varsigma(s^c), \breve{\varsigma}(\upsilon)\right) - V^0_{C_3} & \text{if } i = *. \end{cases}$$

9. Update the label of υ:

$$\varsigma(\upsilon) := \begin{cases} \breve{\varsigma}(\upsilon) & \text{if } \log \tau \leq -\frac{\Delta U}{T}, \\ \varsigma(\upsilon) & \text{otherwise.} \end{cases}$$

where τ is a constant threshold ($\tau \in (0, 1)$).

10. If $j < |\mathcal{V}|$: $\{j := j + 1$ and goto step 3.$\}$
11. Set $T := T_{k+1}$, $k := k + 1$, $j := 1$ and goto step 3, until convergence (i.e. the number of the changed labels between the k^{th} and $(k+1)^{\text{th}}$ iteration is lower than a threshold.)
Note: in the tests, we used $\tau = 0.3$, $T_0 = 4$, and an exponential heating strategy: $T_{k+1} = 0.96 \cdot T_k$

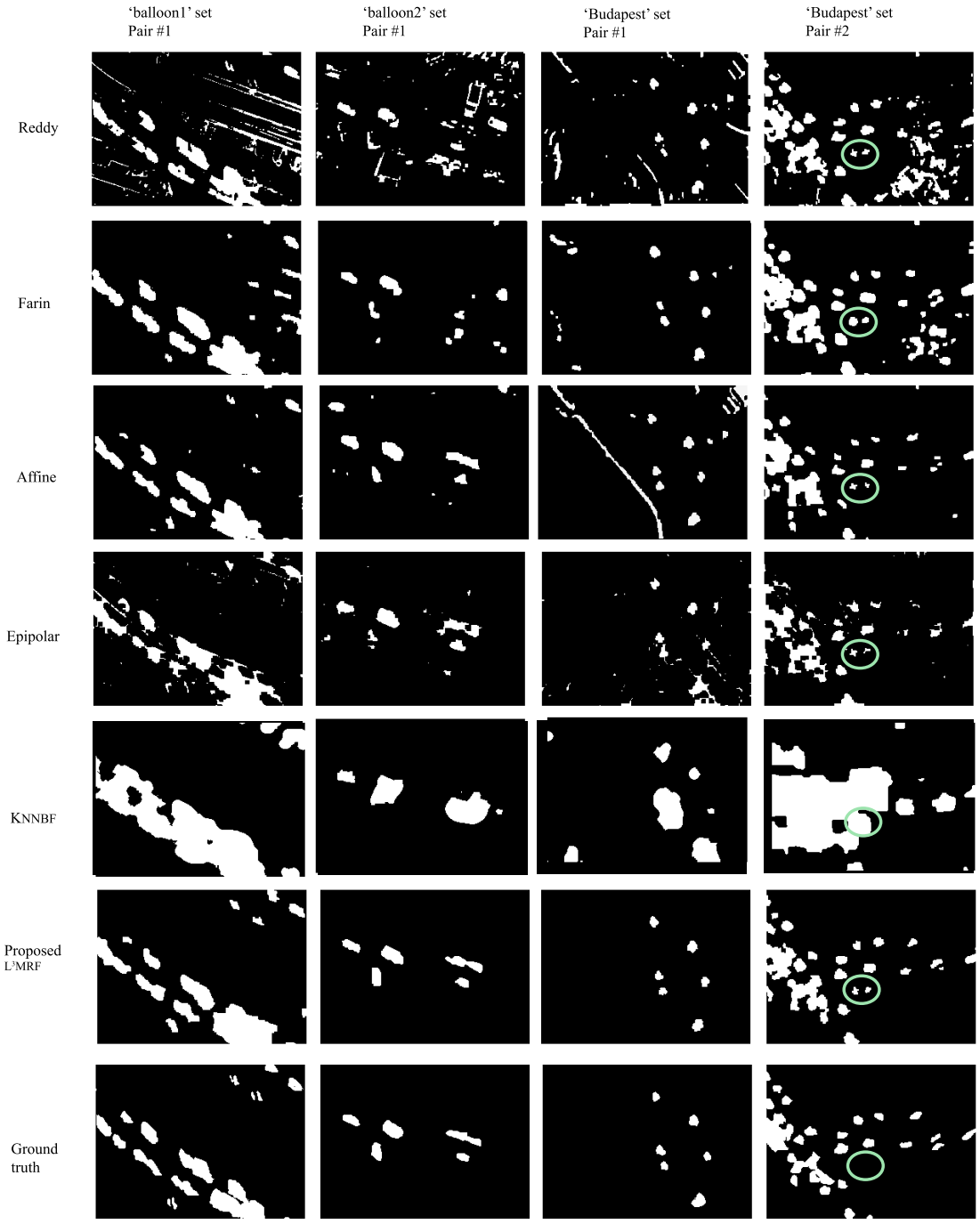


Figure B.1: Comparative segmentations with different test methods and Ground Truth using the image pairs of Fig. 3.4. In the right column, the ellipses demonstrate a limitation: a high standing lamp is detected as a false moving object by all methods.

dc_942_14

Appendix C

Supplement regarding Multitemporal Marked Point Processes

This appendix provides supplementary materials regarding the proposed object level change detection and target sequence analysis methodologies.

C.1 Object level change detection

This section supports the presentation of the object level change detection approach described in Sec. 4.2.

Table C.1 contains quantitative evaluation results of building change detection, with comparing the proposed mMPP method to the Post Detection Classification (PDC) approach [145] .

Algorithm C.1 provides a detailed pseudo code of the iterative Bi-layer Multiple Birth and Death (bMBD) technique developed for optimizing the energy function of the multitemporal Marked Process (mMPP) model.

Fig. C.1 shows detection results with the proposed mMPP approach and various reference techniques (described in Sec. 4.2.4) for qualitative comparison.

Table C.1: Quantitative evaluation results of building change detection.

			FN		FP		MC		FC		Pix. F-sc.	
Data Set	#CH	#UC	PDC	**mMP**	PDC	**mMP**	PDC	**mMP**	PDC	**mMP**	PDC	**mMP**
BUDAPEST	20	21	3	0	7	2	1	0	9	2	0.72	0.78
BEIJING	13	4	1	0	2	1	0	0	3	0	0.77	0.85
SZADA	50	7	4	2	0	1	3	4	3	0	0.76	0.82
ABIDJAN	0	21	2	0	2	0	0	0	4	0	0.78	0.91

#CH and #UC denote the total number of changed resp. unchanged buildings in the set. PDC denotes the Post Detection Classification reference method and mMP refers to the proposed multitemporal Marked Point Process model. Evaluation rates FN, FP, MC, FC and DA are introduced in Sec. 4.2.4.

Algorithm C.1: Bi-layer Multiple Birth and Death (bMBD) optimization

1. Initialization: calculate the $P_b^{(i)}(s)$ ($i \in \{1,2\}$) and $P_{\text{ch}}(s)$ birth maps, and start with an empty population $\omega = \emptyset$. Since the main goal of the *combined birth map* in each image is to keep focus on all building candidate areas, we derive it with the maximum operator from the birth maps of the features. For example, when gradient, color and shadow are simultaneously used, we obtain the final field as $P_b(s) = \max\left\{P_b^{\text{gr}}(s), P_b^{\text{co}}(s), P_b^{\text{sh}}(s)\right\}\ \forall s \in S$.

2. Main program: initialize the inverse temperature parameter $\beta = \beta_0$ and the discretization step $\delta = \delta_0$ and alternate birth and death steps:

   - *Birth step*: for each pixel $s \in S$, if there is no object with center s in the current configuration ω, pick up $\xi \in \{1, 2, *\}$ randomly, let be

   $$\widehat{P}_b = \begin{cases} P_{\text{ch}}(s) \cdot P_b^{(\xi)}(s) & \text{if } \xi \in \{1,2\} \\ \left(1 - P_{\text{ch}}(s)\right) \cdot \max\left\{P_b^{(1)}(s), P_b^{(2)}(s)\right\} & \text{if } \xi = * \end{cases}$$

   and execute the following birth process with probability $\delta\widehat{P}_b$:

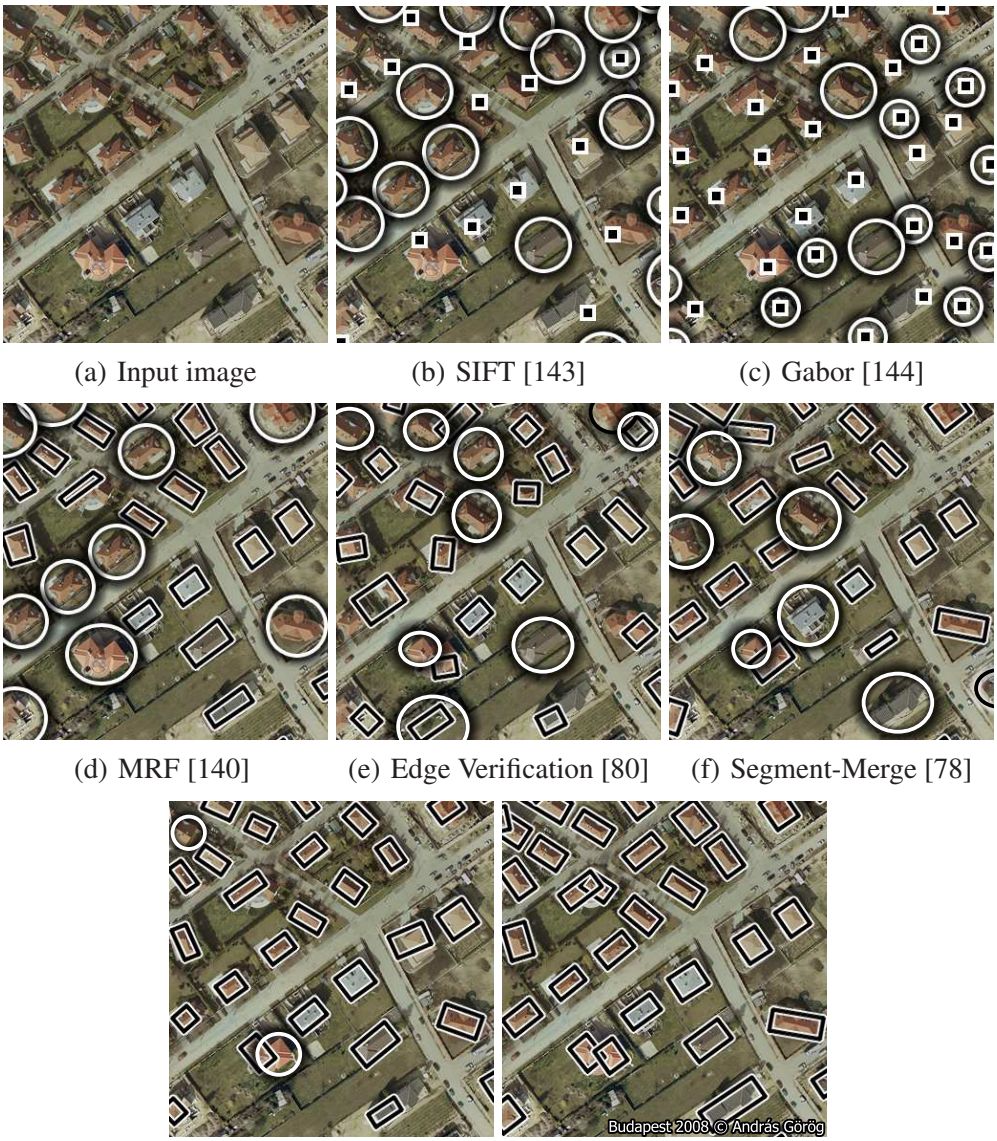
     - generate a new object u with center s and image index ξ
     - set the orientation $\theta(u)$ following the $\eta(., \mu_\theta^{(\xi)}(s), \sigma_\theta)$ Gaussian distribution as shown in Sec. 4.2.2.1
     - add u to the current configuration ω

   - *Death step*: Consider the configuration of objects $\omega = \{u_1, \ldots, u_n\}$ and sort it from the highest to the lowest value of $A_{\mathcal{F}}(u)$. For each object u taken in this order, compute $\Delta\Phi_\omega(u) = \Phi_{\mathcal{F}}(\omega/\{u\}) - \Phi_{\mathcal{F}}(\omega)$, derive the *death rate* as follows:

   $$d_\omega(u) = \frac{\delta a_\omega(u)}{1 + \delta a_\omega(u)}, \quad \text{with} \quad a_\omega(u) = e^{-\beta \cdot \Delta\Phi_\omega(u)}$$

   and remove u from ω with probability $d_\omega(u)$. Note that according to eq. (4.2), $\Delta\Phi_\omega(u)$ depends only on u and its neighbours in ω, thus $d_\omega(u)$ can be calculated locally without computing the global configuration energies $\Phi_{\mathcal{F}}(\omega/\{u\})$ and $\Phi_{\mathcal{F}}(\omega)$.

   - *Convergence test*: if the process has not converged, increase the inverse temperature β and decrease the discretization step δ by a geometric scheme and go back to the birth step. Convergence is obtained when all the objects added during the birth step, and only these ones, have been killed during the death step.



(a) Input image (b) SIFT [143] (c) Gabor [144]

(d) MRF [140] (e) Edge Verification [80] (f) Segment-Merge [78]

(g) Proposed MPP (h) Ground Truth

Figure C.1: Evaluation of the single view building model. Comparing the proposed MPP model to the SIFT [143], Gabor [144], MRF [140], Edge Verification (EV) [80], Segment-Merge (SM) [78] methods, and to the Ground Truth. Circles denote completely missing or false objects. SIFT and Gabor only extract building centers.

C.2 A point process model for target sequence analysis

This section details specific algorithms developed for Inverse Synthetic Aperture Radar (ISAR) image based target sequence analysis, with additional tables and figures supporting the evaluation process.

First, a Markov Random Field based robust foreground-background segmentation algorithm is described for the noisy ISAR frames. *Second*, we introduce the object generation kernels and the pseudo code of the optimization algorithm. *Third*, we provide additional tables to support quantitative evaluation presented in Sec. 4.3.6.

C.2.1 Foreground-background separation of ISAR frames

This section describes a Markov Random Field (MRF) approach, which we proposed in [7] for segmenting the ISAR images into *foreground* and *background* classes, with decreasing the spurious effects of speckle noise.

The goal is to obtain a binary label map $\varpi = \{\upsilon(s)|s \in S\}$, where $\upsilon_s \in \{\mathrm{fg}, \mathrm{bg}\}$ labels correspond to the foreground and background classes, respectively. Assuming that the ISAR amplitude values in both classes follow log-normal distributions [149], we model the $p_{\mathrm{bg}}(s) = P(g_t(s)|\upsilon(s) = \mathrm{bg})$ and $p_{\mathrm{fg}}(s) = P(g_t(s)|\upsilon(s) = \mathrm{fg})$ log-amplitude posterior probabilities by Gaussian densities.

To estimate the Gaussian distribution parameters we used a semi-supervised approach. We could assume having a prior estimation about the ratio of foreground areas compared to the image size, since our vessel targets have shown a typical line segment structure, and the imaging step has intended to provide us spatially normalized images where the target is centered and the image is cropped so that it estimates a narrow bounding box of the target. Using this ratio, the upper part of the image histogram has formed the training regions for foreground, the lower one for background.

Let us denote by $\mathbf{1}_s^{\mathrm{fg}} \in \{0, 1\}$ the indicator function of the foreground class in a given segmentation, where $\mathbf{1}_s^{\mathrm{fg}} = 1$ iff $\upsilon(s) = \mathrm{fg}$. We denote by $s \sim r$, if pixel s is in the 4-neighborhood of pixel r in the S lattice. The optimal foreground mask is derived through minimizing the the following MRF energy [194] function:

$$\varpi_{\mathrm{opt}} = \underset{\varpi \in 2^S}{\operatorname{argmin}} \sum_{s \in S} \log \frac{p_{\mathrm{fg}}(s)}{p_{\mathrm{bg}}(s)} \cdot \mathbf{1}_s^{\mathrm{fg}} + \sum_{r \sim s} \beta \left(\mathbf{1}_s^{\mathrm{fg}} \cdot \mathbf{1}_r^{\mathrm{fg}} + (1 - \mathbf{1}_s^{\mathrm{fg}}) \cdot (1 - \mathbf{1}_r^{\mathrm{fg}}) \right) \tag{C.1}$$

Since (C.1) belongs to the F^2 class of energy functions [194], efficient graph cut based optimization [105] can provide the optimal B mask, as demonstrated in Fig. C.2.

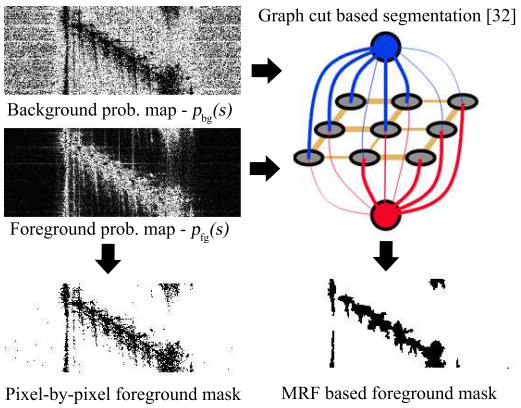


Figure C.2: Demonstration of the foreground-background segmentation. *Top left:* background and foreground probability maps (high probabilities indicated with greater intensities), *bottom left:* foreground mask through pixel-by-pixel maximum likelihood classification (only for reference), *top right:* sketch of graph-cut based MRF optimization [105], *bottom right:* foreground mask (B) by the proposed MRF model

C.2.2 F^{m}MPP energy optimization

For supporting the discussion of F^{m}MPP energy optimization algorithm in Sec. 4.3.5, we provide here the pseudo-code of object candidate generation (Algorithm C.2) and the steps of iterative energy minimization (Algorithm C.3).

C.2.3 Quantitative evaluation of the F^{m}MPP method

Table C.2 summarizes the axis level detection rates (smaller error values are favored) for the three steps of the workflow: (a) *Initial* detection, (b) *RANSAC*-based refinement and (c) the final F^{m}MPP output after iterative optimization. We can observe that the errors decrease over the consecutive steps, and at the end of the process the summarized E_{AX} rate is between 3% and 7% in all sequences.

By examining the evaluation rates of Table C.2, we can observe that the proposed method can accurately deal with all the seven test cases (`SHIP1-SHIP7`). The improvement between the outputs of the *Initial* and *Optimized F^{m}MPP* phases of the process is particularly significant in the `SHIP1` (shown in Fig. 4.9), `SHIP2` and `SHIP5` sequences, which contain difficult test cases. The developments are also remarkable in the `SHIP3`, `SHIP4` and `SHIP6` cases (see sample frames in Fig. 4.10), while the `SHIP7` sequence contains noisier images with several blurred frames, where the final error rates remain larger (see also the last row of Fig. 4.10).

Algorithm C.2: Object generator function pool

Variables

n: number of frames of the sequence
t: frame index

Notation

$N(\mu, \sigma)$: Gaussian distribution with μ mean value and σ standard deviation.

Functions

Object u = **function** Propose_Obj_by_**PERTURBATION** (t)

- Pick up Δ randomly following the upcoming normal distributions:

  switch(t)

  case 1: $P(\Delta = 0) = 2/3$, $P(\Delta = +1) = 1/3$
  case n: $P(\Delta = 0) = 2/3$, $P(\Delta = -1) = 1/3$
  otherwise: $P(\Delta = -1) = 1/4$, $P(\Delta = 0) = 1/2$,
  $P(\Delta = +1) = 1/4$

  end-switch

- Generate a new object u and set its parameters randomly following the upcoming distributions:

  $x(u) \sim N(x(u_{t+\Delta}), \sigma_x)$, $y(u) \sim N(y(u_{t+i}), \sigma_y)$, $\theta(u) \sim N(\theta(u_{t+\Delta}), \sigma_\theta)$, $l(u) \sim N(l(u_{t+\Delta}), \sigma_l)$

  where σ_x, σ_y, σ_θ and σ_l are model parameters.

- Fill the scatterer vector of u by cloning the scatterer vector from u_t, and randomly add new scatterers from the *pre. scatterer candidate set* of frame $t + \Delta$, or delete some of the actual scatterers

return object u

Object u = **function** Propose_Obj_by_**RANSAC** (t)

- Generate a new object u
- Determine the axis line of u by applying RANSAC to the scatterer candidates of the tth frame.
- Estimate the endpoints of the axis line segment by morphology.
- Determine the scatterers for u by the preliminary **Scatterer Filtering (SF) Kernel**.

return object u

Algorithm C.3: Pseudo code of the multiframe energy optimization

Variables

n: number of frames of the sequence

k: iteration counter

Steps of the algorithm

1) Initialize the configuration with the output of the deterministic detector: $\omega^{[0]} = \{u_1^{[0]}, u_2^{[0]}, \ldots, u_n^{[0]}\}$, and set iteration counter $k = 0$, inverse temperature $\beta = \beta_0$, refinement parameter $\delta = \delta_0$ and boolean **STOP**:=`false`.

2) Iterate the following steps while **STOP**=`false`.

for each $t = 1, \ldots, n$:

- Pick up $\Psi_{\text{Birth}} \in \{\text{PERT}, \text{RANSAC}\}$ randomly
- Generate a new object u so that:

  if $\Psi_{\text{Birth}} = \text{PERT}$:
  $u :=$ Propose_Obj_by_**PERTURBATION**(t)
  if $\Psi_{\text{Birth}} = \text{RANSAC}$:
  $u :=$ Propose_Obj_by_**RANSAC**(t)
- Consider the ω^* configuration which could be obtained if in $\omega^{[k]}$ we exchanged $u_t^{[k]}$ by u.
- Calculate the energy difference between and $\omega^{[k]}$ and ω^*:

$$\Delta\Phi_\omega(u,t) = \Phi_{\mathcal{F}}(\omega^*) - \Phi_{\mathcal{F}}\left(\omega^{[k]}\right)$$

- Calculate the $d_\omega(u)$ exchange rate as follows:

$$d_\omega(u) = \frac{\delta a_\omega(u)}{1 + \delta a_\omega(u)} \text{ with } a_\omega(u) = e^{-\beta \cdot \Delta\Phi_\omega(u)}$$

and set

$$u_t^{[k+1]} = \begin{cases} u & \text{with probability } d_\omega(u) \\ u_t^{[k]} & \text{otherwise} \end{cases}$$

3) $k := k + 1$, increase β and decrease δ with a geometric scheme.

4) If the process converged: **STOP**:=`true`.

Table C.2: Evaluation of the different steps in the F^{m}MPP model for the test sequences.

Sequence	*Step*	Axis extraction errors					Scatterer level errors			
		E^x_{AX}	E^y_{AX}	E^l_{AX}	E^θ_{AX}	$\mathbf{E}_{\mathrm{AX}}$	TP	FP	FN	E_{SP}
SHIP1	*Initial*	6.31	9.89	10.6	5.64	*23.6*	249	117	111	*7.4*
	RANSAC	5.11	5.69	9.11	2.18	*15.3*	339	49	21	*2.2*
	F^{m}MPP	**0.44**	**0.27**	**3.73**	**0.8**	**3.8**	**349**	**10**	**11**	**0.5**
SHIP2	*Initial*	5.85	1.72	13.11	1.51	*12.3*	680	49	40	*4.8*
	RANSAC	2.99	1.02	6.56	0.71	*6.2*	703	23	17	*1.6*
	F^{m}MPP	**0.47**	**0.17**	**4.29**	**0.58**	**3.2**	**718**	**2**	**2**	**0.4**
SHIP3	*Initial*	2.80	2.15	5.70	2.15	*10.3*	301	33	19	*1.5*
	RANSAC	1.65	1.33	4.92	1.52	*7.5*	306	30	14	*1.6*
	F^{m}MPP	0.33	0.30	2.65	0.90	**3.4**	**311**	**22**	**9**	**1.0**
SHIP4	*Initial*	2.37	0.83	5.96	0.58	*5.7*	696	66	24	*1.1*
	RANSAC	2.70	0.82	5.69	0.79	*6.0*	699	64	21	*1.0*
	F^{m}MPP	**0.64**	**0.06**	**4.37**	**0.38**	**3.2**	**705**	**22**	**15**	**0.7**
SHIP5	*Initial*	2.07	0.96	5.86	1.10	*6.4*	691	69	29	*0.9*
	RANSAC	1.43	0.47	3.50	0.86	*4.1*	695	71	25	*0.6*
	F^{m}MPP	**0.19**	**0.09**	**4.01**	**0.80**	**3.3**	**707**	**29**	**13**	**0.3**
SHIP6	*Initial*	2.33	1.54	3.71	1.96	*7.8*	763	48	47	*0.9*
	RANSAC	1.46	0.70	4.09	1.11	*5.9*	763	49	47	*0.8*
	F^{m}MPP	**0.01**	**0.07**	**3.20**	**0.50**	**3.0**	**764**	**18**	**46**	**0.7**
SHIP7	*Initial*	4.53	0.87	9.27	1.12	*9.9*	562	61	38	*3.5*
	RANSAC	3.32	0.72	9.21	0.75	*8.6*	567	58	33	*2.9*
	F^{m}MPP	**2.13**	**0.13**	**8.13**	**0.56**	**6.7**	**559**	**37**	**41**	**2.5**
AIRPL	*Initial*	1.68	6.16	16.32	2.56	*34.9*	n.a.	n.a.	n.a.	n.a.
	F^{m}MPP	**0.24**	**0.80**	**3.28**	**0.76**	**6.6**	n.a.	n.a.	n.a.	n.a.

Axis detection error rates are the following: E^x_{AX}, E^y_{AX}, E^l_{AX} and E^θ_{AX} mean parameter errors are measured in pixels, the normalized E_{AX} error rate is expressed in percent (%). Regarding the Scatterer Detection, the number of False Positive (FP) and False Negative (FN) scatterers determine the precision and recall factors of the process, while the E_{SP} rate shows the scatterer positioning accuracy (low values are preferred)

Appendix D

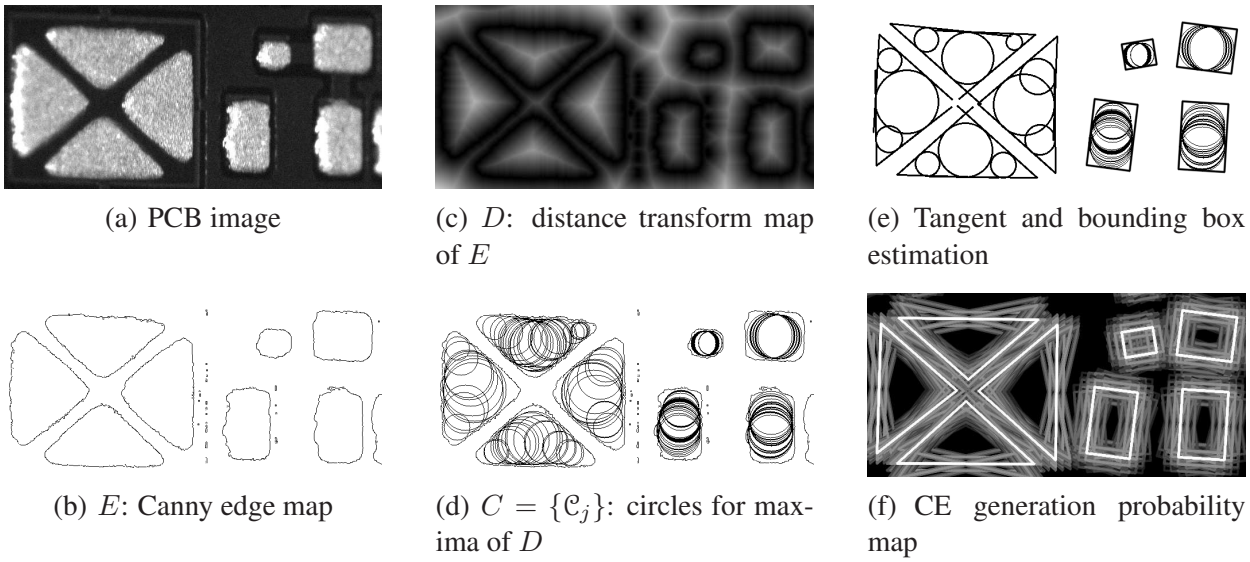
Supplement regarding Embedded Marked Point Processes

This appendix offers supplementary materials regarding Embedded Marked Point Processes.

First, details are provided regarding the Bottom-Up Stochastic Entity Proposal (BUSEP) process in PCB analysis. As introduced in Sec. 5.4 instead of applying fully random sampling, we construct a data driven stochastic entity generation scheme, which proposes relevant parent objects with higher probability based on various image features. Here we give a concrete implementation of the BUSEP algorithm, developed for the printed circuit board (PCB) analysis application. In this case we have to deal with variously shaped and scaled circuit elements (CEs): rectangles, ellipses and triangles, while the size of CEs can be notably different (see Fig. D.1(a)), which factors significantly increase the size of the parameter space. We use in the preprocessing step a binary *foreground* mask B obtained by Otsu's thresholding method from the input image, which realizes a coarse separation of the circuit entities (i.e. foreground) from the board (i.e. background). However, due to notable noise, this B mask can be unreliable for purposes of CE separation and shape estimation. In addition, some neighboring CEs may also be merged into one blob in the mask. The process starts with CE candidate generation based on the *foreground* mask, as described in Algorithm D.1. Thereafter, using these CE candidates probabilistic parameter maps (i.e. extended birth maps) are calculated for the BUSEP process, which is detailed in Algorithm D.2. Finally, the parent object generation step - which uses the parameter maps - is realized by Algorithm D.3.

Second, we give the pseudo code of the Multi-level Multiple Birth and Death algorithm in Algorithm D.4. This description refers to the proposed generic algorithm, however, the parent object generation step is naturally application depended, for example, the PCB analysis task can use Algorithm D.3.

Third, regarding the built in area analysis application we give here two additional figures: Fig. D.2 which demonstrates the efficiency of (child level) chimney detection, and Fig. D.3 for qualitative comparison of the sMPP and EMPP approaches discussed in Sec. 5.7.



(a) PCB image

(c) D: distance transform map of E

(e) Tangent and bounding box estimation

(b) E: Canny edge map

(d) $C = \{\mathcal{C}_j\}$: circles for maxima of D

(f) CE generation probability map

Figure D.1: Steps of the bottom-up entity proposal process

Algorithm D.1: CE Candidate Generation

Step 1. Generate the Canny edge map E of the PCB image (Fig. D.1(b))

Step 2. Generate the distance transform map of E, and denote it by D (Fig. D.1(c))

Step 3. Find *l*ocal ma*x*ima pixels in D: $\{s^i_{\text{lx}} | i = 1, 2, \ldots, n_{\text{lx}}\}$, and for each i draw a $\mathcal{C}$ circle with center s^i_{lx} and radius $D(s^i_{\text{lx}})$. Keep only circles which correspond in majority to foreground regions of the coarse B *foreground* mask: $C = \{\mathcal{C}_j | j = 1, \ldots, n_c\}$. We denote henceforward by $\mathcal{C}_j \overset{\text{int}}{\sim} \mathcal{C}_i$ if $\mathcal{C}_j$ and $\mathcal{C}_i$ circles intersect (Fig. D.1(d)).

Step 4. We define an indirect intersection relation $\overset{\text{i}C'}{\sim}$ for a subset $C' \subset C$ where for each $\mathcal{C}_j, \mathcal{C}_i \in C'$: $\mathcal{C}_j \overset{\text{i}C'}{\sim} \mathcal{C}_i$ iff $\mathcal{C}_j \overset{\text{int}}{\sim} \mathcal{C}_i$ or $\exists \mathcal{C}_k \in C'$: $\mathcal{C}_j \overset{\text{int}}{\sim} \mathcal{C}_k$ AND $\mathcal{C}_k \overset{\text{i}C'}{\sim} \mathcal{C}_i$

Step 5. We prepare an n-partition of $C = C_1 \cup C_2 \cup \ldots \cup C_n$ so that for each l each $\mathcal{C}_j \in C_l$ is in $\overset{\text{i}C_l}{\sim}$ relation with all elements in C_l, but not with any other circles from $C \backslash C_l$ (see in Fig. D.1(d) the grouped circles).

Step 6. To all partitions obtained above we assign a CE candidate. For each C_l we calculate the radius-variation of the included circles. If the variation if high enough we mark the object as a *triangle* candidate, otherwise as a *R&E* (rectangle or ellipse) candidate.

Algorithm D.2: BUSEP parameter map calculation

Step 1. Generate a binary *foreground* mask B obtained by Otsu's thresholding method from the input PCB image, which realizes a coarse separation of the circuit entities (i.e. foreground) from the board (i.e. background).

Step 2. Call Algorithm D.1 for CE Candidate Generation.

Step 3. Deal separately with the *R&E* (rectangle or ellipse) and the *Triangle candidates* in the following ways:

R&E candidates: for each object, estimate the bounding rectangle $\mathcal{R}$ of the union of the corresponding circles (Fig. D.1(e), right). Denote the R&E object candidates as: $\{\mathcal{R}_1, ...\mathcal{R}_{n_r}\}$ and let $o(\mathcal{R}_i)$ be the center of $\mathcal{R}_i$. Then, for each pixel s, we determine the closest rectangle $\mathcal{R}_s^{\min} = \text{argmin}_i \, ||s - o(\mathcal{R}_i)||$ and calculate the birth value:

$$P_b^{\mathcal{R}}(s) = k_{\mathcal{R}}\left(\frac{||s - o(\mathcal{R}_s^{\min})||}{h_{\mathcal{R}}}\right) \tag{D.1}$$

with a $k_{\mathcal{R}}(.)$ kernel function, and $h_{\mathcal{R}}$ bandwidth parameter [10]. Besides marking the candidate regions of the rectangular or elliptical CE centers, the $\{\mathcal{R}_i | i = 1 \dots n_r\}$ set provides local estimations for the side/axis length and orientation parameters: $\mu_M^{\mathcal{R}}(s) = a_M(\mathcal{R}_s^{\min})$, $\mu_m^{\mathcal{R}}(s) = a_m(\mathcal{R}_s^{\min})$ and $\mu_\theta^{\mathcal{R}}(s) = \theta(\mathcal{R}_s^{\min})$.

Triangle candidates: determine the circles with the minimal and maximal radius of the group, and the circle which has the highest distance from the minimal circle (Fig. D.1(e), left part). Calculate joint tangents of the maximal and minimal circles. Estimate the triangle sides accordingly. Let us assume that we have detected n_t triangle candidates: $\{\mathcal{T}_1, ...\mathcal{T}_{n_t}\}$, and similarly to the R&E case, we derive here a triangle birth map $P_b^{\mathcal{T}}(.)$ with estimated side length and orientation values $\mu_M^{\mathcal{T}}(.)$, $\mu_m^{\mathcal{T}}(.)$ and $\mu_\theta^{\mathcal{T}}(.)$.

Step 4. Let be the summarized birth value $\widehat{P}_b(s) = P_b^{\mathcal{R}}(s) + P_b^{\mathcal{T}}(s) + P_0$ for each pixel s.

Algorithm D.3: Parent_object_generation(input: pixel s)

Step 1. Execute one from the following three options:

(a) with a probability $P_b^{\mathcal{R}}(s)/\widehat{P}_b(s)$ generate a rectangle or ellipse patent object, u with center $o(u) := s$. Set the side lengths/axes and orientation parameters as $a_M(u) = \mu_M^{\mathcal{R}}(s) + \eta_M$, $a_m(u) = \mu_m^{\mathcal{R}}(s) + \eta_m$ and $\theta(u) = \mu_\theta^{\mathcal{R}}(s) + \eta_\theta$, where η_M, η_m and η_θ are independent zero mean Gaussian random variables.

(b) with a probability $P_b^{\mathcal{T}}(s)/\widehat{P}_b(s)$ generate a triangle u with reference point $o(u) := s$. Set the geometric parameters based on the $\mu_{M,m,\theta}^{\mathcal{T}}$ maps, similarly to the previous case.

(c) otherwise, generate arbitrary typed CE object u with reference point s, and set its geometric parameters fully randomly following prior size distributions.

Step 2. Initialize u without any children: $Q_u = \text{nil}$

Algorithm D.4: Multi-level Multiple Birth and Death algorithm

1) Initialization: start with empty population $\omega = \emptyset$, set the birth rate b_0, initialize the inverse temperature parameter $\beta = \beta_0$ and the discretization step $\delta = \delta_0$. Calculate the BUSEP parameter maps according to *Algorithm D.2*.

2) Main program: alternate the following three steps:

- *Birth step*: Visit all pixels on the image lattice S one after another. At each pixel s, call Parent_object_generation(s) of *Algorithm D.3* with probability $\delta \cdot \widehat{P}_b(s)$. For each new object u, with a probability $p_u^0 = \mathbf{1}_{\omega=\emptyset} + \mathbf{1}_{\omega\neq\emptyset} \cdot \min_{\psi_j \in \omega} \hat{d}_{\psi_j}(u)$, generate a new ψ empty segment (i.e. object group), add u to ψ and ψ to ω. Otherwise, add u to an existing segment $\psi_i \in \omega$ with a probability $p_u^i = (1 - \hat{d}_{\psi_i}(u))/\sum_{\psi_j \in \omega}(1 - \hat{d}_{\psi_j}(u))$.
- *Death step*: Consider the actual configuration of all objects within ω and sort it by decreasing values depending on $A(u) + I_g(u, \psi)\big|_{u\in\psi}$. For each object u taken in this order, compute $\Delta\Phi_\omega(u) = \Phi_{\mathcal{F}}(\omega/\{u\}) - \Phi_{\mathcal{F}}(\omega)$, derive the *death rate* $p_\omega^d(u)$ as

$$p_\omega^d(u) = \Gamma(\Delta\Phi_\omega(u)) = \frac{\delta \exp(-\beta \cdot \Delta\Phi_\omega(u))}{1 + \delta \exp(-\beta \cdot \Delta\Phi_\omega(u))}, \tag{D.2}$$

  and delete object u with probability $p_\omega^d(u)$. Remove empty population segments from ω, if they appear.
- *Group re-arrangement*: Consider the objects of the current ω population, one after another. For each object u of segment ψ we propose an alternative object u', so that the shape type of u', may be different from u, and the geometric parameters of u' are derived from the parameters of u by adding zero mean Gaussian random values. The next step is selecting a group candidate for u'. For this reason, we randomly choose a v object from the proximity neighborhood of u ($v \in \mathcal{N}_u(\omega)$), and assign u' to the group of v, denoted by ψ'. Then, we estimate the energy cost of exchanging $u \in \psi$ to $u' \in \psi'$:

$$\Delta\varphi(\omega, u, u') = \varphi_Y(u') - \varphi_Y(u) + \sum_{v \prec \omega\setminus\{u\}} \big[I_p(u', v) - I_p(u, v)\big] + I_g(u', \psi') - I_g(u, \psi)$$

  The *object exchange rate* is calculated using the $\Gamma(.)$ function defined by (D.2):

$$p_\omega^{\mathrm{e}}(u, u') = \Gamma\Big(\Delta\varphi(\omega, u, u')\Big)$$

  Finally with a probability $p_\omega^{\mathrm{e}}(u, u')$, we replace u with u'.
- *Child Maintenance* For each $u \prec \omega$ object:
  - add new child objects to Q_u randomly
  - sort Q_u by decreasing values depending on the $\varphi_d^c(u, q_u)$ values
  - for each child object $q_u \in Q_u$ taken in this order, compute the child removal rate $d_u^c(q_u)$ similarly to the parent level, but considering only the child level unary and interaction terms.
  - remove q_u from Q_u with a probability $d_u^c(q_u)$

3) Convergence test: if the process has not converged yet, increase β and decrease δ with a geometric scheme, and go back to the birth step.

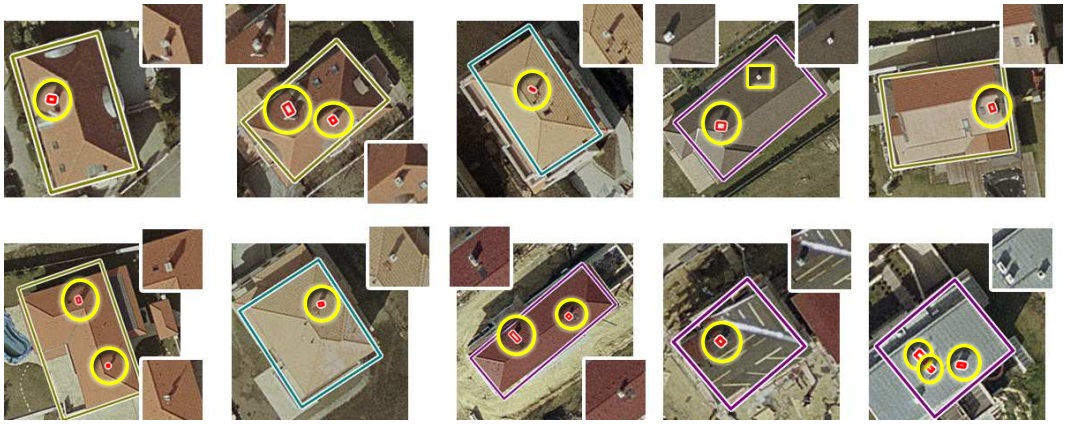
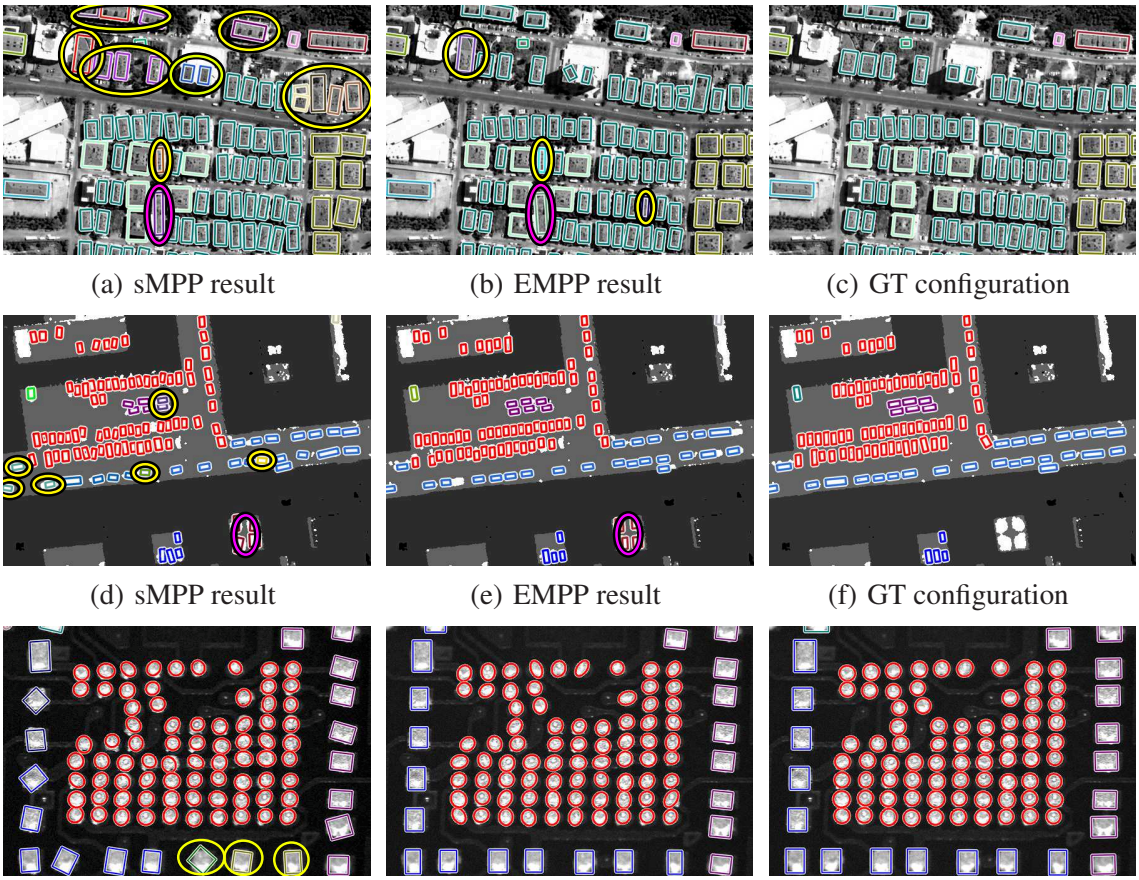


Figure D.2: Building analysis - sample results for chimney detection. True hits are marked by yellow circles, a false negative is highlighted in the third image of the upper row by a yellow rectangle. In the corners of the samples, the raw images of the chimney regions are displayed separately for visual verification



(a) sMPP result (b) EMPP result (c) GT configuration

(d) sMPP result (e) EMPP result (f) GT configuration

(g) sMPP result (h) EMPP result (i) GT configuration

Figure D.3: Qualitative validation of the sMPP and the EMPP configurations versus the Ground Truth (only parent a group levels are displayed). Yellow ellipses mark grouping errors and purple ones false objects. By building analysis (row 1) groups of houses and condos are separated

dc_942_14

Appendix E

Supplement regarding 4D environment perception

This appendix provides supplementary materials regarding our contributions in 4D environment perception from Chapter 6.

Algorithm E.1 presents the details of our fast bounding box estimation method for partially extracted object blobs from Lidar data, introduced in Sec. 6.4.2.

Algorithm E.2 provides the pseudo code of the Hough transform based point cloud frame alignment algorithm discussed in Sec. 6.4.5, while Fig. E.1 shows an example for multimodal point cloud registration with our proposed approach.

Table E.1 shows quantitative evaluation results regarding the object separation algorithm (Sec. 6.4.2).

We evaluated the proposed multimodal registration process with matching the measurements of Velodyne sensors to the MLS point clouds. Quantitative analysis result of the matching process is given in Table E.2. Since *Ground Truth* transformation was not available, we calculated first the asymmetric Modified Hausdorff Distance (MHD) between the $\mathcal{P}_{\text{RMB}}$ Velodyne and $\mathcal{P}_{\text{MLS}}$ MLS clouds:

$$\text{MHD}(\mathcal{P}_{\text{RMB}}, \mathcal{P}_{\text{MLS}}) = \frac{1}{\#\mathcal{P}_{\text{RMB}}} \sum_{p \in \mathcal{P}_{\text{RMB}}} \min_{q \in \mathcal{P}_{\text{MLS}}} dist(p, q)$$

where $\#\mathcal{P}$ denotes set cardinality. Columns 5-7 of Table E.2 contain the obtained MHD values initially, after the object based Hough matching step, and in the final stage following NDT-based registration refinement. We can observe that both steps significantly decrease the distances between the scans in almost all data sets. However, the absolute MHD values do not reflect properly the accuracy of the algorithm, since the presence of several moving objects, especially large trams or tracks, mislead (increase) the calculated average distances. For this reason, we also used a modified error metrics called Median Point Distance (MPD), where we sort the points in $\mathcal{P}_{\text{RMB}}$ from the lowest to the highest value of $\min_q dist(p, q)$, and take the median of the distances among all

$p \in \mathcal{P}_{\text{RMB}}$. As shown in the 8-10th columns of Table E.2 the MPD values are also significantly decreased during the registration process, and in seven out of the eight scenes the resulting MPD errors are below 3cm, which fact was also confirmed by visual verification. Only the test scene Bajcsy yielded erroneous registration result both by visual and quantitative (MHD, MPD) analysis. In this sample both RMB point clouds contained several moving vehicles, including large buses which occluded various relevant scene structure elements. The 11th (last) column of Table E.2 lists for each scene the computational time of the complete matching process (varying between 0.3 and 2.2 seconds), which confirms that the approach is close to online usability.

Algorithm E.1: Fast bounding box estimation for partially detected objects

Input: 2D ground projection of the objects extracted at the coarse level of the two-level grid based algorithm

Step 1. Estimate the boundary cells of the objects, and construct the convex hull from the boundary using the monotone chain algorithm [195]. Let be $i = 1$.

Step 2. Visit the consecutive point pairs of the hull p_i and p_{i+1}, one after another ($i = 1, 2, \ldots, i_{\max}$):

- Consider the line l_i between point p_i and p_{i+1}, as a side candidate of the bounding box rectangle.
- Find the $p_\star$ point of the hull, whose distance is maximal from l_i, and draw a $l_\star$ parallel line with l_i which intersects $p_\star$. We consider $l_\star$ as the second side candidate of the bounding box.
- Project all the points of the convex hull to the line l_i, and find the two extreme ones p' and p''. The remaining two sides of the bounding box candidate will be constructed by taking perpendicular lines to l_i, which intersect p' and p'' respectively.

Step 3. Chose the optimal bounding box from the above generated rectangle set by minimizing the average distance between the points of the convex hull and the fitted rectangle.

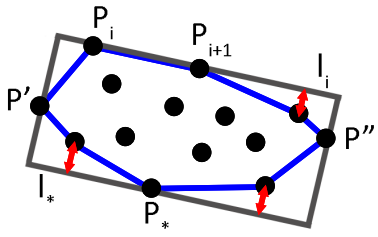


Figure D.1 Demonstration of the fast 2D bounding box fitting algorithm for the convex hull of the top-view object projection (the bounding box is shown marked by gray color)

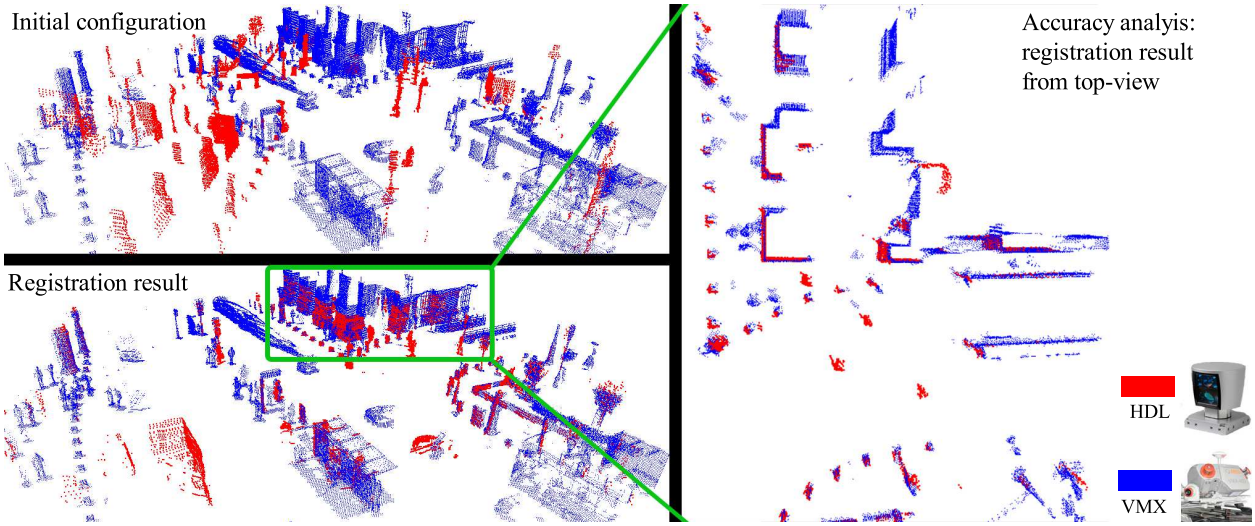


Figure E.1: Multimodal Velodyne HDL-64E to Riegl VMX-450 registration results using our proposed method (Fővám tér, Budapest).

Algorithm E.2: The cloud alignment algorithm. Takes two clouds as inputs and calculates the transformation between them. $Rot(\alpha)$ represents the rotational matrix along z axis.

procedure SCANALIGNMENT($F1$, $F2$, T)
  $C1 \leftarrow ObjectDetect(F1)$
  $C2 \leftarrow ObjectDetect(F2)$
  Initialize 3D accumulator A
  for all $o1 \in C1$ **do**
    for all $o2 \in C2$ **do**
      for $\alpha \in [0, 359]$ **do**
        $o1' \leftarrow Rot(\alpha) * o1$
        $(dx, dy) \leftarrow o2 - o1'$
        $A[dx, dy, \alpha] \leftarrow A[dx, dy, \alpha] + 1$
      end for
    end for
  end for
  $\alpha, dx, dy \leftarrow FindMaximum(A)$
  $F1, T1 \leftarrow TransformCloud(F1, \alpha, dx, dy)$
  $F1, T2 \leftarrow NDT(F1, F2)$
  $T \leftarrow T2 * T1$
end procedure

Table E.1: Numerical comparison of the detection results obtained by the Connected Component Analysis [188] and the proposed *Hierarchal Grid Model*. The number of objects (NO) are listed for each data set, and also in aggregate.

Point Cloud Dataset	NO	Conn. Comp. Analysis [188]		Prop. Hierarchical Grid	
		F-rate(%)	Avg. Processing Speed (fps)	F-rate(%)	Avg. Processing Speed (fps)
Budapest Dataset #1	669	77	0.38	89	29
Budapest Dataset #2	429	64	0.22	79	25
KITTI Dataset [196]	496	75	0.46	82	29
Overall	1594	72	0.35	83	28

Table E.2: Results of multimodal RMB Lidar and MLS point cloud registration (Velodyne HDL-64E/VLP-16 to Riegl-VMX scan matching)

Scene	Sensor	initial offset, rotation	MHD (m)			MPD (m)			Comput
			Init	Hough	Final†	Init	Hough	Final	time
Astoria hub	HDL	2.2m, 62°	3.641	0.773	0.415	1.587	0.511	0.022	1.923
	VLP	2.2m, 99°	5.045	0.582	0.221	3.623	0.231	0.008	0.665
Bajcsy road	HDL	2.0m, 92°	5.657	11.441	10.105	1.177	2.702	4.539	0.992
	VLP	10.3m, 72°	6.971	20.115	17.796	4.179	17.319	14.341	0.329
Deák square	HDL	1.4m, 32°	3.638	0.717	0.338	1.516	0.345	0.004	1.960
	VLP	3.6m, 127°	7.348	0.870	0.911	5.502	0.143	0.101	0.769
Fővám square	HDL	2.0m, 134°	8.404	3.494	2.870	6.143	1.339	0.008	3.796
	VLP	0.1m, 20°	5.143	1.849	1.431	3.393	0.216	0.010	1.182
Kálvin square1	HDL	1.4m, 118°	9.891	0.774	0.205	5.808	0.469	0.005	1.159
	VLP	2.0m, 42°	11.427	7.016	8.178	5.007	0.752	0.014	0.573
Kálvin square2	HDL	5.8m, 104°	19.445	2.252	2.002	4.968	0.437	0.023	0.288
	VLP	6.1m, 56°	19.663	2.901	5.909	16.826	0.817	0.065	0.221
Múzeum boulevard	HDL	2.2m, 70°	14.911	3.358	1.373	12.354	1.315	0.009	2.574
	VLP	5.0m, 91°	6.970	2.489	3.412	1.477	0.312	0.018	1.403
Gellért square	HDL	1.0m, 125°	3.180	0.949	1.046	1.238	0.224	0.014	1.045
	VLP	0.0m, 34°	5.241	2.438	1.574	4.037	1.173	0.029	0.852
Average values‡	HDL	2.3m, 92°	9.016	1.760	1.178	4.802	0.663	0.012	1.821
	VLP	3.7m, 68°	8.691	2.592	3.091	5.695	0.521	0.035	0.809

Error measures: MHD: Modified Hausdorff distance, MPD: median point distance.
†Final result refers to the Hough+NDT cascade, ‡Bajcsy was excluded from averaging, due to unsuccessful registration

References

The author's SCI journal publications

[1] **C. Benedek**, B. Gálai, B. Nagy, and Z. Jankó, "Lidar-based gait analysis and activity recognition in a 4D surveillance system," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 1, pp. 101–113, 2018. IF: 3.558*. 6.3.6.2, 7.2

[2] **C. Benedek**, "An embedded marked point process framework for three-level object population analysis," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4430–4445, 2017. IF: 5.071. 5.1, 7.2

[3] A. Börcs, B. Nagy, and **C. Benedek**, "Instant object detection in Lidar point clouds," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 7, pp. 992 – 996, 2017. IF: 2.892. 6.4, 6.4.3, 6.4(a), 7.2

[4] **C. Benedek**, M. Shadaydeh, Z. Kato, T. Szirányi, and J. Zerubia, "Multilayer Markov random field models for change detection in optical remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 107, pp. 22–37, 2015. Special Issue on Multitemporal Remote Sensing Change Detection, IF: 4.188. 3.3.3, 7.2

[5] A. Börcs and **C. Benedek**, "Extraction of vehicle groups in airborne lidar point clouds with two-level point processes," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 3, pp. 1475–1489, 2015. IF: 3.360. 5.5.2, 5.5.2, 5.6, 5.7, 5.7, 7.2, 7.2

[6] **C. Benedek**, "3D people surveillance on range data sequences of a rotating Lidar," *Pattern Recognition Letters*, vol. 50, pp. 149–158, 2014. Special Issue on Depth Image Analysis, IF: 1.551. 6.3.1, 6.3.2, 7.2

[7] **C. Benedek** and M. Martorella, "Moving target analysis in ISAR image sequences with a multiframe Marked Point Process model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 4, pp. 2234–2246, 2014. IF: 3.514. 4.3.3, 4.3.4, 5.1, 7.2, 7.3, C.2.1

[8] **C. Benedek**, O. Krammer, M. Janóczki, and L. Jakab, "Solder paste scooping detection by multi-level visual inspection of printed circuit boards," *IEEE Transactions on Industrial Electronics*, vol. 60, no. 6, 2013. IF: 6.500. 5.1, 5.4, 5.5.3, 5.5.3, 5.6, 5.7, 7.2, 7.2

[9] Á. Utasi and **C. Benedek**, "A Bayesian approach on people localization in multi-camera systems," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 1, pp. 105–115, 2013. IF: 2.259. 1, 4.2.4, 6.2.4, 7.2

[10] **C. Benedek**, X. Descombes, and J. Zerubia, "Building development monitoring in multitemporal remotely sensed image pairs with stochastic birth-death dynamics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 33–50, 2012. IF: 4.795. 2.5, 2.7, 4.2.2.2, 4.2.4, 5.5.1, 5.6, 5.7, 7.2, D

[11] **C. Benedek**, "Detection of soldering defects in printed circuit boards with hierarchical Marked Point Processes," *Pattern Recognition Letters*, vol. 32, no. 13, pp. 1535 – 1543, 2011. IF: 1.034. 5.1, 5.5.3, 5.7, 5.4, 7.2

[12] **C. Benedek**, T. Szirányi, Z. Kato, and J. Zerubia, "Detection of object motion regions in aerial image pairs with a multi-layer Markovian model," *IEEE Transactions on Image Processing*, vol. 18, no. 10, pp. 2303–2315, 2009. IF: 2.848. 3.1, 3.2.4, 3.3, 3.3.2, 7.2

[13] **C. Benedek** and T. Szirányi, "Change detection in optical aerial images by a multi-layer conditional mixed Markov model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 10, pp. 3416–3430, 2009. IF: 2.234. 3.1, 3.3.3, 4.2.1, 4.3.1, 7.2

[14] **C. Benedek** and T. Szirányi, "Bayesian foreground and shadow detection in uncertain frame rate surveillance videos," *IEEE Transactions on Image Processing*, vol. 17, no. 4, pp. 608–621, 2008. IF: 3.315. 5.5.1, 6.1, 6.3, 6.2.1, 1

[15] **C. Benedek** and T. Szirányi, "Study on color space selection for detecting cast shadows in video surveillance," *International Journal of Imaging Systems and Technology*, vol. 17, no. 3, pp. 190–201, 2007. Special Issue on Applied Color Image Processing, IF: 0.482. 1, 6.2.1

The author's international conference publications[1]

[16] B. Nagy and **C. Benedek**, "Real-time point cloud alignment for vehicle localization in a high resolution 3D map," in *Workshop on Computer Vision for Road Scene Understanding and Autonomous Driving at ECCV'18*, vol. 11129 of *Lecture Notes in Computer Science*, pp. 226–239, Munich, Germany: Springer, 2019. 6.4, 7.2

[1]Conference papers published in proceedings, books or in non-SCI periodicals

[17] B. Gálai and **C. Benedek**, "Change detection in urban streets by a real time Lidar scanner and MLS reference data," in *International Conference on Image Analysis and Recognition (ICIAR)*, vol. 10317 of *Lecture Notes in Computer Science*, pp. 210–220, Montral, Canada: Springer, July 2017. 6.4, 6.4.6, 6.4(b), 7.2

[18] B. Nagy and **C. Benedek**, "3D CNN based phantom object removing from mobile laser scanning data," in *International Joint Conference on Neural Networks (IJCNN)*, (Anchorage, Alaska, USA), pp. 4429–4435, May 2017. 6.4, 6.4.4, 7.2

[19] B. Gálai and **C. Benedek**, "Gait recognition with compact lidar sensors," in *International Conference on Computer Vision Theory and Applications (VISAPP)*, (Porto, Portugal), pp. 1–7, Feb 2017. 6.3.3, 7.2

[20] B. Gálai, B. Nagy, and **C. Benedek**, "Crossmodal point cloud registration in the Hough space for mobile laser scanning data," in *International Conference on Pattern Recognition (ICPR)*, (Cancun, Mexico), pp. 3363–3368, IEEE, Dec 2016. 6.4, 6.4.5, 6.4.7, 7.2

[21] B. Gálai and **C. Benedek**, "Feature selection for lidar-based gait recognition," in *International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM)*, (Prague, Czech Republic), pp. 1–5, IEEE, Oct 2015. 6.3.6.2, 7.2

[22] **C. Benedek**, B. Nagy, B. Gálai, and Z. Jankó, "Lidar-based gait analysis in people tracking and 4D visualization," in *European Signal Processing Conference (EUSIPCO)*, (Nice, France), pp. 1138–1142, Aug 2015. 6.3.3, 6.3.3, 7.2

[23] A. Börcs, B. Nagy, M. Baticz, and **C. Benedek**, "A model-based approach for fast vehicle detection in continuously streamed urban LIDAR point clouds," in *Workshop on Scene Understanding for Autonomous Systems at ACCV'14*, vol. 9008 of *Lecture Notes in Computer Science*, pp. 413–425, Singapore: Springer, April 2015. 6.4, 6.4.2, 7.2

[24] A. Börcs, B. Nagy, and **C. Benedek**, "Fast 3-D urban object detection on streaming point clouds," in *Workshop on Computer Vision for Road Scene Understanding and Autonomous Driving at ECCV'14*, vol. 8926 of *Lecture Notes in Computer Science*, pp. 628–639, Zürich, Switzerland: Springer, March 2015. 6.4, 6.4.2, 6.4.7, 7.2

[25] **C. Benedek**, "Hierarchical image content analysis with an embedded marked point process framework," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Florence, Italy), pp. 5147–5151, May 2014. 5.1, 7.2

[26] A. Börcs, B. Nagy, and **C. Benedek**, "On board 3D object perception in dynamic urban scenes," in *IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, (Budapest, Hungary), pp. 515–520, December 2013. 6.4, 7.3

[27] **C. Benedek**, Z. Jankó, C. Horváth, D. Molnár, D. Chetverikov, and T. Szirányi, "An integrated 4D vision and visualisation system," in *International Conference on Computer Vision Systems (ICVS)*, vol. 7963 of *Lecture Notes in Computer Science*, pp. 21–30, St. Petersburg, Russia: Springer, July 2013. 6.1, 7.3

[28] **C. Benedek**, "A two-layer marked point process framework for multilevel object population analysis," in *International Conference on Image Analysis and Recognition (ICIAR)*, vol. 7950 of *Lecture Notes in Computer Science*, pp. 160–169, Póvoa de Varzim, Portugal: Springer, June 2013. 5.3, 7.2

[29] A. Börcs, O. Józsa, and **C. Benedek**, "Object extraction in urban environments from large-scale dynamic point cloud dataset," in *IEEE International Workshop on Content-Based Multimedia Indexing (CBMI)*, (Veszprém, Hungary), pp. 191–194, June 2013. 6.4, 7.3

[30] O. Józsa, A. Börcs, and **C. Benedek**, "Towards 4D virtual city reconstruction from Lidar point cloud sequences," in *ISPRS Workshop on 3D Virtual City Modeling*, vol. II-3/W1 of *ISPRS Annals Photogram. Rem. Sens. and Spat. Inf. Sci.*, pp. 15–20, May 2013. 6.4, 6.4.1, 6.4.5, 7.2

[31] **C. Benedek**, D. Molnár, and T. Szirányi, "A dynamic MRF model for foreground detection on range data sequences of rotating multi-beam Lidar," in *International Workshop on Depth Image Analysis (WDIA)*, vol. 7854 of *Lecture Notes in Computer Science*, pp. 87–96, Tsukuba City, Japan: Springer, November 2012. 6.3.1, 7.2

[32] A. Börcs and **C. Benedek**, "Urban traffic monitoring from aerial LIDAR data with a two-level marked point process model," in *International Conference on Pattern Recognition (ICPR)*, (Tsukuba City, Japan), pp. 1379–1382, November 2012. 5.7, 7.2

[33] **C. Benedek** and M. Martorella, "Ship structure extraction in ISAR image sequences by a Markovian approach," in *IET International Conference on Radar Systems*, (Glasgow, UK), October 2012. 4.3.3, 7.2

[34] A. Börcs and **C. Benedek**, "A marked point process model for vehicle detection in aerial LIDAR point clouds," in *XXII. ISPRS Congress*, vol. I-3 of *ISPRS Annals Photogram. Rem. Sens. and Spat. Inf. Sci.*, pp. 93–98, August 2012. 2.7, 7.2

[35] Á. Utasi and **C. Benedek**, "A multi-view annotation tool for people detection evaluation," in *Proceedings of the First International Workshop on Visual Interfaces for Ground Truth Collection in Computer Vision Applications*, (Capri, Italy), May 2012. 6.2.4, 7.2

[36] D. Baltieri, R. Vezzani, R. Cucchiara, Á. Utasi, **C. Benedek**, and T. Szirányi, "Multi-view people surveillance using 3D information," in *Proceedings of The 11th International Workshop on Visual Surveillance*, (Barcelona, Spain), pp. 1817–1824, November 2011. 6.2, 7.2

[37] **C. Benedek**, "Analysis of solder paste scooping with hierarchical point processes," in *Proc. IEEE International Conference on Image Processing (ICIP)*, (Brussels, Belgium), pp. 2121–2124, September 2011. 5.5.3, 7.2

[38] **C. Benedek** and M. Martorella, "ISAR image sequence based automatic target recognition by using a multi-frame marked point process model," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, (Vancouver, Canada), pp. 3791–3794, July 2011. 4.3.1, 7.2

[39] Á. Utasi and **C. Benedek**, "A 3-D marked point process model for multi-view people detection," in *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, (Colorado Springs, USA), pp. 3385–3392, June 2011. 6.2, 7.2

[40] A. Kovács, **C. Benedek**, and T. Szirányi, "A joint approach of building localization and outline extraction," in *IASTED International Conference on Signal Processing, Pattern Recognition and Applications (SPPRA)*, (Innsbruck, Austria), February 2011. 4.2.1, 7.2

[41] Á. Utasi and **C. Benedek**, "Multi-camera people localization and height estimation using multiple birth-and-death dynamics," in *Workshop on Visual Surveillance*, (Queenstown, NZ), November 2010. 6.2, 7.2

[42] **C. Benedek**, X. Descombes, and J. Zerubia, "Building detection in a single remotely sensed image with a point process of rectangles," in *International Conference on Pattern Recognition (ICPR)*, (Istanbul, Turkey), August 2010. 4.2.1, 7.2

[43] **C. Benedek**, "Efficient building change detection in sparsely populated areas using coupled marked point processes," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, (Honolulu, Hawaii, USA), July 2010. 4.2.1, 7.2

[44] **C. Benedek**, X. Descombes, and J. Zerubia, "Building extraction and change detection in multitemporal remotely sensed images with multiple birth and death dynamics," in *IEEE Workshop on Applications of Computer Vision (WACV)*, (Snowbird, Utah, USA), pp. 100–105, December 2009. 4.2.1, 7.2

[45] **C. Benedek** and T. Szirányi, "A mixed markov model for change detection in aerial photos with large time differences," in *Proc. International Conference on Pattern Recognition (ICPR)*, (Tampa, FL, USA), IAPR, December 2008. 3.3, 7.2

[46] **C. Benedek**, T. Szirányi, Z. Kato, and J. Zerubia, "A multi-layer MRF model for object-motion detection in unregistered airborne image-pairs," in *Proc. IEEE International Conference on Image Processing (ICIP)*, vol. VI, (San Antonio, Texas, USA), pp. 141–144, IEEE, September 2007. 3.2, 7.2

[47] **C. Benedek** and T. Szirányi, "Markovian framework for foreground-background-shadow separation of real world video scenes," in *Asian Conference on Computer Vision (ACCV)*, vol. 3851 of *Lecture Notes in Computer Science*, pp. 898–907, Hyderabad, India: Springer, January 2006. 6.3

The author's other connected publications

[48] **C. Benedek**, Z. Jankó, T. Szirányi, D. Chetverikov, O. Józsa, A. Börcs, and I. Eichardt, "Method and system for generating a three-dimensional model." US Patent No. 10,163,256 (PCT number: PCT–HU2014–000017), Dec. 2018. 7.1, 7.3

[49] **C. Benedek**, "Probabilistic Approaches on Environment Perception with 2D-3D Imaging Sensors." Habilitation thesis, Péter Pázmány Catholic University, Faculty of Information Technology and Bionics, June 2017. 7.4

[50] A. Börcs, B. Nagy, and **C. Benedek**, "Dynamic environment perception and 4D reconstruction using a mobile rotating multi-beam Lidar sensor," in *Handling Uncertainty and Networked Structure in Robot Control*, Studies in Systems, Decision and Control, pp. 153–180, Springer, 2016. 6.1, 7.2

[51] **C. Benedek** and T. Szirányi, "Shadow detection in digital images and video," in *Computational Photography: Methods and Applications*, Digital Imaging and Computer Vision, pp. 283–312, CRC Press, Taylor & Francis, 2010. 6.2.1, 1

[52] **C. Benedek**, X. Descombes, and J. Zerubia, "Building extraction and change detection in multitemporal aerial and satellite images in a joint stochastic approach," Research Report 7143, INRIA, Sophia Antipolis, December 2009. 4.2.2

[53] **C. Benedek**, *Novel Markovian Change Detection Models in Computer Vision*. PhD thesis, Péter Pázmány Catholic University, Faculty of Information Technology, June 2008. 2.3, 2.1.4, 1, 7.2, 7.2

[54] **C. Benedek**, T. Szirányi, Z. Kato, and J. Zerubia, "A three-layer MRF model for object motion detection in airborne images," Research Report 6208, INRIA Sophia Antipolis, France, June 2007. 3.2, 7.2

The author's national conference publications

[55] B. Nagy and **C. Benedek**, "3D CNN alapú MLS pontfelhőszegmentáció (3D CNN based MLS point cloud segmentation)," in *Conference of Hungarian Association for Image Analysis and Pattern Recognititon (KÉPAF)*, (Debrecen, Hungary), January 2019. In Hungarian. 7.4

[56] Ö. Zováthi, B. Nagy, and **C. Benedek**, "Valós idejű pontfelhőillesztés és járműlokalizáció nagy felbontású 3D térképen (Real time point cloud alignment and vehicle localization in high resolution 3D map)," in *Conference of Hungarian Association for Image Analysis and Pattern Recognititon (KÉPAF)*, (Debrecen, Hungary), January 2019. In Hungarian. 7.4

[57] B. Gálai and **C. Benedek**, "Járás alapú személyazonosítás és cselekvésfelismerés LiDAR szenzorokkal (Gait based person identification and action recognition with LiDAR sensors)," in *Conference of Hungarian Association for Image Analysis and Pattern Recognititon (KÉPAF)*, (Sovata, Romania), January 2017. In Hungarian. 7.4

[58] B. Nagy, B. Gálai, and **C. Benedek**, "Multimodális pontfelhőregisztráció Hough tér alapú előillesztéssel (Multimodal point cloud registration with Hough based initial matching)," in *Conference of Hungarian Association for Image Analysis and Pattern Recognititon (KÉPAF)*, (Sovata, Romania), January 2017. In Hungarian. 7.4

[59] A. Börcs, B. Nagy, and **C. Benedek**, "Utcai objektumok gyors osztályozása LIDAR pontfelhősorozatokon (fast classification of urban objects in Lidar point cloud sequences," in *Conference of Hungarian Association for Image Analysis and Pattern Recognititon (KÉPAF)*, (Sovata, Romania), January 2017. In Hungarian. 7.4

[60] **C. Benedek**, "Hierarchikus jelölt pontfolyamat modell objektumpopulációk többszintű elemzéséhez (Hierarchical marked point process model for multi-level object population analysis)," in *Conference of Hungarian Association for Image Analysis and Pattern Recognititon (KÉPAF)*, (Kerekegyháza, Hungary), January 2015. In Hungarian. 7.4

[61] A. Börcs, B. Nagy, and **C. Benedek**, "Valós idejű járműdetekció LIDAR pontfelhő sorozatokon (Real time vehicle detection in LIDAR point cloud sequences)," in *Conference of Hungarian Association for Image Analysis and Pattern Recognititon (KÉPAF)*, (Kerekegyháza, Hungary), January 2015. In Hungarian, received Attila Kuba Prize. 7.4

[62] B. Nagy, **C. Benedek**, and Z. Jankó, "Mozgó személyek követése és 4D vizualizációja Lidar alapú járáselemzéssel (Moving people tracking and 4D visualization through Lidar-based gait analysis)," in *Conference of Hungarian Association for Image Analysis and Pattern Recognititon (KÉPAF)*, (Kerekegyháza, Hungary), January 2015. In Hungarian. 7.4

[63] P. Polcz and **C. Benedek**, "3D mesh generation from aerial LiDAR point cloud data," in *Hungarian Computer Graphics and Geometry Conference (GRAFGEO)*, (Budapest, Hungary), 2014. 7.4

[64] **C. Benedek**, Z. Jankó, A. Börcs, I. Eichhardt, D. Chetverikov, and T. Szirányi, "Viewpoint-free video synthesis with an integrated 4D system," in *Hungarian Computer Graphics and Geometry Conference (GRAFGEO)*, (Budapest, Hungary), 2014. 7.4

[65] **C. Benedek** and M. Martorella, "Mozgó célpontok vizsgálata radar képsorozatokon jelölt pontfolyamat modellell (Moving target analysis in radar image sequences with marked point processes)," in *Conference of Hungarian Association for Image Analysis and Pattern Recognititon (KÉPAF)*, (Bakonybél, Hungary), January 2013. In Hungarian. 7.4

[66] A. Börcs and **C. Benedek**, "Városi forgalomfelügyelet kétszintű jelölt pontfolyamat modellel légi LiDAR felvételeken (Urban traffic control from aerial LIDAR data with a two-layer marked point process model)," in *Conference of Hungarian Association for Image Analysis and Pattern Recognititon (KÉPAF)*, (Bakonybél, Hungary), January 2013. In Hungarian. 7.4

[67] C. Horváth, D. Molnár, **C. Benedek**, and T. Szirányi, "MRF alapú előtérdetekció és objektumkövetés lidar pontfelhő szekvenciákon (MRF based foreground detection and multi target tracking in LIDAR point cloud sequences)," in *Conference of Hungarian Association for Image Analysis and Pattern Recognititon (KÉPAF)*, (Bakonybél, Hungary), January 2013. In Hungarian. 7.4

[68] Á. Utasi and **C. Benedek**, "Személyek lokalizálása és magasságuk becslése többszörös születés és halál dinamikával többkamerás környezetben (Multi-camera people localization and height estimation using multiple birth-and-death dynamics)," in *Conference of Hungarian Association for Image Analysis and Pattern Recognititon (KÉPAF)*, (Szeged, Hungary), January 2011. In Hungarian. 7.4

[69] **C. Benedek**, X. Descombes, and J. Zerubia, "Épületek és változásaik detekciója sztochasztikus módszerekkel (Detecting buildings and their changes with a stochastic approach)," in *Conference of Hungarian Association for Image Analysis and Pattern Recognititon (KÉPAF)*, (Szeged, Hungary), January 2011. In Hungarian. 7.4

[70] **C. Benedek** and T. Szirányi, "Kevert markov modell alapú változásdetekció nagy időkülönbséggel készült légi fotókon (Mixed Markovian Change Detection in Aerial Images with Large Time Differences)," in *Conference of Hungarian Association for Image Analysis and Pattern Recognititon (KÉPAF)*, (Budapest, Hungary), January 2009. In Hungarian. 7.4

Publications connected to the dissertation

[71] G. Scarpa, R. Gaetano, M. Haindl, and J. Zerubia, "Hierarchical multiple Markov chain model for unsupervised texture segmentation," *IEEE Trans. on Image Processing*, vol. 18, no. 8, pp. 1830–1843, 2009. 1, 7.2

[72] J. Porway, Q. Wang, and S. C. Zhu, "A hierarchical and contextual model for aerial image parsing," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 254–283, 2010. 1, 5.1, 7.2

[73] Z. Kato and J. Zerubia, *Markov random fields in image segmentation*. Foundations and Trends in Signal Processing, Now Publishers, 2012. 1, 2.1.4

[74] D. Benboudjema and W. Pieczynski, "Unsupervised statistical segmentation of nonstationary images using triplet Markov fields," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 8, pp. 1367–1378, 2007. 1

[75] A. Fridman, "Mixed Markov models.," *Proc. National Academy of Sciences of USA*, vol. 100, no. 14, pp. 8092–8096, 2003. 1, 2.1.5, 2.1.5, 3.3, 3.3.1, 3.3.2, 3.5, 7.1, 7.2

[76] S. Kumar and M. Hebert, "Discriminative random fields," *International Journal of Computer Vision*, vol. 68, no. 2, pp. 179–202, 2006. 1

[77] S. Z. Li, *Markov random field modeling in computer vision*. London, UK: Springer-Verlag, 1995. 1, 2.1.4

[78] S. Müller and D. Zaum, "Robust building detection in aerial images," in *ISPRS Object Extraction for 3D City Models, Road Databases and Traffic Monitoring - Concepts, Algorithms and Evaluation, (CMRT05)*, (Vienna, Austria), pp. 143–148, 2005. 1, 4.2.4, C.1(f), C.1

[79] P. Soille, *Morphological Image Analysis: Principles and Applications*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2 ed., 2003. 1

[80] B. Sirmaçek and C. Ünsalan, "Building detection from aerial imagery using invariant color features and shadow information," in *Int. Symp. on Computer and Information Sciences (ISCIS)*, (Istanbul, Turkey), 2008. 1, 4.2.2.1, 4.2.4, C.1(e), C.1

[81] Z. Song, C. Pan, Q. Yang, F. Li, and W. Li, "Building roof detection from a single high-resolution satellite image in dense urban area," in *ISPRS Congress*, (Beijing, China), pp. 271–277, 2008. 1

[82] F. Lafarge, X. Descombes, J. Zerubia, and M. Pierrot-Deseilligny, "Structural approach for building reconstruction from a single DSM," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 135–147, 2010. 1, 5.5.1

[83] X. Descombes and J. Zerubia, "Marked point processes in image analysis," *IEEE Signal Processing Magazine*, vol. 19, no. 5, pp. 77–84, 2002. 1, 7.1

[84] T. Blaskovics, Z. Kato, and I. Jermyn, "A Markov random field model for extracting near-circular shapes," in *International Conference on Image Processing (ICIP)*, (Cairo, Egypt), pp. 1073–1076, 2009. 1

[85] C. Wang, F. Liao, and C. Ma, "Detection of pedestrian crossing from focus to spread," in *World Congress on Intelligent Control and Automation (WCICA)*, pp. 4897–4901, 2012. 1

[86] X. Descombes, ed., *Stochastic geometry for image analysis*. Digital Signal and Image Processing, Wiley-ISTE, 2011. 1

[87] X. Descombes, R. Minlos, and E. Zhizhina, "Object extraction using a stochastic birth-and-death dynamics in continuum," *Journal of Mathematical Imaging and Vision*, vol. 33, pp. 347–359, 2009. 1, 2.2.3, 2.7, 4.3.5, 4.4, 5.4, 5.5.3

[88] A. Gamal-Eldin, X. Descombes, and J. Zerubia, "Multiple birth and cut algorithm for point process optimization," in *International Conference on Signal-Image Technology and Internet-Based Systems (SITIS)*, (Kuala Lumpur, Malaysia), pp. 35–42, 2010. 1

[89] Y. Verdie and F. Lafarge, "Detecting parametric objects in large scenes by Monte Carlo sampling," *International Journal of Computer Vision*, vol. 106, pp. 57–75, 2014. 1, 2.2.3

[90] S. Ben Hadj, F. Chatelain, X. Descombes, and J. Zerubia, "Parameter estimation for a marked point process within a framework of multidimensional shape extraction from remote sensing images," in *ISPRS Technical Commission III Symposium on Photogrammetry Computer Vision and Image Analysis (PCV)*, (Paris, France), 2010. 1

[91] F. Chatelain, X. Descombes, and J. Zerubia, "Parameter estimation for marked point processes. application to object extraction from remote sensing images," in *Energy Minimization Methods in Comp. Vision and Pattern Recogn.*, vol. 5681 of *Lecture Notes in Computer Science*, pp. 221–234, Bonn, Germany: Springer, 2009. 1

[92] Y. Li and J. Li, "Oil spill detection from SAR intensity imagery using a marked point process," *Remote Sensing of Environment*, vol. 114, no. 7, pp. 1590 – 1601, 2010. 1

[93] F. Chatelain, A. Costard, and O. J. J. Michel, "A Bayesian marked point process for object detection. Application to muse hyperspectral data," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Prague, Czech Republic), pp. 3628–3631, 2011. 1

[94] A. Gamal-Eldin, X. Descombes, and J. Zerubia, "A novel algorithm for occlusions and perspective effects using a 3D object process," in *IEEE International Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, (Prague, Czech Republic), pp. 1569–1572, 2011. 1

[95] C. Mallet, F. Lafarge, M. Roux, U. Soergel, F. Bretar, and C. Heipke, "A marked point process for modeling Lidar waveforms," *IEEE Trans. on Image Processing*, vol. 19, no. 12, pp. 3204–3221, 2010. 1

[96] B.-T. Vo and B.-N. Vo, "Labeled random finite sets and multi-object conjugate priors," *IEEE Trans. on Signal Processing*, vol. 61, no. 13, pp. 3460–3475, 2013. 1

[97] M. Bredif, O. Tournaire, B. Vallet, and N. Champion, "Extracting polygonal building footprints from digital surface models: A fully-automatic global optimization framework," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 77, no. 1, pp. 57–65, 2013. 1

[98] J. Zhou, C. Proisy, P. Couteron, X. Descombes, J. Zerubia, G. le Maire, and Y. Nouvellon, "Tree crown detection in high resolution optical images during the early growth stages of eucalyptus plantations in brazil," in *Asian Conf. on Pattern Rec.*, pp. 623–627, 2011. 1

[99] Y. Yu, J. Li, H. Guan, C. Wang, and M. Cheng, "A marked point process for automated tree detection from mobile laser scanning point cloud data," in *International Conference on Computer Vision in Remote Sensing (CVRS)*, (Xiamen, China), pp. 140–145, IEEE, 2012. 1

[100] A. Veillard, S. Bressan, and D. Racoceanu, "SVM-based framework for the robust extraction of objects from histopathological images using color, texture, scale and geometry," in *International Conference on Machine Learning and Applications (ICMLA)*, vol. 1, pp. 70–75, 2012. 1

[101] W. Ge and R. Collins, "Marked point processes for crowd counting," in *IEEE Conference on Computer Vision and Pattern Recognition*, (Miami, FL, USA), pp. 2913–2920, 2009. 1

[102] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721–741, 1984. 2, 2, 2.1.1, 2.1.4, 3.1, 3.2.3, 3.3.2, 3.5, 7.1

[103] E. Aarts and J. Korst, *Simulated Annealing and Boltzman Machines*. New York: John Wiley & Sons, 1990. 2, 2.1.4, 2.4

[104] R. Potts, "Some generalized order-disorder transformation," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 48, no. 1, pp. 106–109, 1952. 2, 2.1.3

[105] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1124–1137, 2004. 2.1.4, 6.3.1.3, 6.3.6.1, 6.4.6, C.2.1, C.2

[106] Z. Kato, *Modélisations markoviennes multirésolutions en vision par ordinateur. Application ŕ la segmentation d'images SPOT (Multiresolution Markovian models in computer vision. Application on segmentation of SPOT images)*. PhD thesis, University of Nice, INRIA, Sophia Antipolis, France, 1994. Available in French and English. 2.1.4

[107] Z. Kato, J. Zerubia, and M. Berthod, "Satellite image classification using a modified Metropolis dynamics," in *International Conference on Acoustics, Speech and Signal Processing*, pp. 573–576, March 1992. 2.1.4, 3.2.3, B

[108] M. Ortner, X. Descombes, and J. Zerubia, "A marked point process of rectangles and segments for automatic analysis of digital elevation models.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 1, pp. 105–119, 2008. 2.2.3, 2.4

[109] F. Lafarge, G. Gimel'farb, and X. Descombes, "Geometric feature extraction by a multi-marked point process," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1597 –1609, 2010. 2.2.3, 5.1, 5.7

[110] Z. Tu and S.-C. Zhu, "Image segmentation by Data-Driven Markov Chain Monte Carlo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, pp. 657–673, 2002. 2.4, 5.4

[111] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org. 2.3

[112] Z. Zhao, H. Li, R. Zhao, and X. Wang, "Crossing-line crowd counting with two-phase deep neural networks," in *European Conference on Computer Vision (ECCV)*, vol. 9912 of *Lecture Notes in Computer Science*, pp. 712–726, Amsterdam, The Netherlands: Springer, 2016. 2.3

[113] R. U. W. Luo, B. Yang, "Fast and Furious: Real Time End-to-End 3D Detection, Tracking and Motion Forecasting with a Single Convolutional Net," in *IEEE Conference on Computer Vision and Pattern Recognition*, (Salt Lake City, Utah, United States), 2018. 2.3

[114] D. Clausi and H. Deng, "Design-based texture feature fusion using Gabor filters and co-occurrence probabilities," *IEEE Trans. on Image Processing*, vol. 14, pp. 925–936, July 2005. 3.1

[115] Z. Kato and T. C. Pong, "A Markov random field image segmentation model for color textured images," *Image and Vision Computing*, vol. 24, no. 10, pp. 1103–1114, 2006. 3.1, 5.1.2

[116] Z. Kato, T. C. Pong, and G. Q. Song, "Multicue MRF image segmentation: Combining texture and color," in *International Conference on Pattern Recognition*, (Quebec, Canada), pp. 660–663, Aug. 2002. 3.1, 3.3.3

[117] T. Szirányi and M. Shadaydeh, "Segmentation of remote sensing images using similarity-measure-based fusion-MRF model," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, pp. 1544–1548, Sept 2014. 3.1, 3.3.3

[118] B. Reddy and B. Chatterji, "An FFT-based technique for translation, rotation and scale-invariant image registration," *IEEE Trans. on Image Processing*, vol. 5, no. 8, pp. 1266–1271, 1996. 3.2, 3.2.4, B

[119] G. J. Hahn and S. S. Shapiro, *Statistical models in engineering*. New York: John Wiley & Sons, 1994, p. 95. 3.2.1, 3.3.1

[120] J. Besag, "On the statistical analysis of dirty images," *Journal of Royal Statistics Society*, vol. 48, pp. 259–302, 1986. 3.2.3

[121] D. Farin and P. de With, "Misregistration errors in change detection algorithms and how to avoid them," in *International Conference on Image Processing (ICIP)*, (Genoa, Italy), pp. 438–441, Sept. 2005. 3.2.4, B

[122] S. Kumar, M. Biswas, and T. Nguyen, "Global motion estimation in spatial and frequency domain," in *IEEE International Conference on Acoustics,Speech,and Signal Processing*, (Montreal, Canada), pp. 333–336, May 2004. 3.2.4

[123] L. Lucchese, "Estimating affine transformations in the frequency domain," in *IEEE International Conference on Image Processing (ICIP)*, vol. II, (Thessaloniki, Greece), pp. 909–912, Sept. 2001. 3.2.4, B

[124] C. Yuan, G. Medioni, J. Kang, and I. Cohen, "Detecting motion regions in the presence of a strong parallax from a moving camera by multiview geometric constraints," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, pp. 1627–1641, September 2007. 3.2.4, B

[125] M. Irani and P. Anandan, "A unified approach to moving object detection in 2D and 3D scenes," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 6, pp. 577–589, 1998. 3.2.4

[126] P. Jodoin and M. Mignotte, "Motion segmentation using a K-nearest-neighbor-based fusion procedure of spatial and temporal label cues," in *Int. Conf. on Image Analysis and Recognition (ICIAR)*, vol. 3656 of *Lecture Notes in Computer Science*, pp. 778–788, Toronto, Canada: Springer, 2005. 3.2.4, B

[127] R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: A systematic survey," *IEEE Trans. on Image Processing*, vol. 14, no. 3, pp. 294–307, 2005. 3.3

[128] J. A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," Tech. Rep. TR-97-021, International Computer Science Institute and Computer Science Division, University of California at Berkley, Berkley, CA, April 1998. 3.3.1

[129] M. Benšić and K. Sabo, "Estimating the width of a uniform distribution when data are measured with additive normal errors with known variance," *Computational Statistics & Data Analysis*, vol. 51, no. 9, pp. 4731–4741, 2007. 3.3.1

[130] R. Wiemker, "An iterative spectral-spatial bayesian labeling approach for unsupervised robust change detection on remotely sensed multispectral imagery," in *Int. Conf. on Computer Analysis of Images and Patterns (CAIP)*, vol. 1296 of *Lecture Notes in Computer Science*, pp. 263–270, Kiel, Germany: Springer, 1997. 3.3.3

[131] S. Ghosh, L. Bruzzone, S. Patra, F. Bovolo, and A. Ghosh, "A context-sensitive technique for unsupervised change detection based on Hopfield-type neural networks," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 45, pp. 778–789, March 2007. 3.3.3

[132] L. Bruzzone and D. Fernández-Prieto, "An adaptive semiparametric and context-based approach to unsupervised change detection in multitemporal remote-sensing images," *IEEE Trans. on Image Processing*, vol. 11, no. 4, pp. 452–466, 2002. 3.3.3

[133] L. Bruzzone, D. Fernandez Prieto, and S. Serpico, "A neural-statistical approach to multitemporal and multisource remote-sensing image classification," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 37, pp. 1350–1359, May 1999. 3.3.3

[134] L. Castellana, A. D'Addabbo, and G. Pasquariello, "A composed supervised/unsupervised approach to improve change detection from remote sensing," *Pattern Recogn. Lett.*, vol. 28, no. 4, pp. 405–413, 2007. 3.3.3

[135] P. Singh, Z. Kato, and J. Zerubia, "A Multilayer Markovian Model for Change Detection in Aerial Image Pairs with Large Time Differences," in *International Conference on Pattern Recognition*, (Stockholm, Sweden), May 2014. 3.3.3

[136] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artif. Intell. Rev.*, vol. 22, no. 2, pp. 85–126, 2004. 3.4

[137] Z. Kato, J. Zerubia, and M. Berthod, "Unsupervised parallel image classification using Markovian models," *Pattern Recognition*, vol. 32, no. 4, pp. 591–604, 1999. 3.4

[138] P. Jodoin, M. Mignotte, and C. Rosenberger, "Segmentation framework based on label field fusion," *IEEE Trans. on Image Processing*, vol. 16, no. 10, pp. 2535–2550, 2007. 3.2.4, 3.5

[139] S. Kumar and M. Hebert, "Detection in natural images using a causal multiscale random field," in *IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR)*, vol. 1, (Madison, USA), pp. 119–126, 2003. 4.2.2.1

[140] A. Katartzis and H. Sahli, "A stochastic framework for the identification of building rooftops using a single remote sensing image," *IEEE Trans. Geosc. Remote Sens.*, vol. 46, no. 1, pp. 259–271, 2008. 4.2.2.1, C.1(d), C.1

[141] V. Tsai, "A comparative study on shadow compensation of color aerial images in invariant color models," *IEEE Trans. Geosc. Remote Sens.*, vol. 44, no. 6, pp. 1661–1671, 2006. 4.2.2.1

[142] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistic Quarterly*, vol. 2, pp. 83–97, 1955. 4.2.4, 4.3.6

[143] B. Sirmaçek and C. Ünsalan, "Urban-area and building detection using SIFT keypoints and graph theory," *IEEE Trans. Geosc. Remote Sens.*, vol. 47, no. 4, pp. 1156–1167, 2009. 4.2.4, C.1(b), C.1

[144] B. Sirmaçek and C. Ünsalan, "A probabilistic framework to detect buildings in aerial and satellite images," *IEEE Trans. Geosc. Remote Sens.*, vol. 49, pp. 211–221, 2011. 4.2.4, C.1(c), C.1

[145] S. Tanathong, K. Rudahl, and S. Goldin, "Object oriented change detection of buildings after the Indian ocean tsunami disaster," in *IEEE International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, (Krabi, Thailand), pp. 65–68, 2008. 4.2.4, C.1

[146] J. L. Walker, "Range-doppler imaging of rotating objects," *IEEE Trans. Aerospace and Electronic Systems*, vol. 16, pp. 23–52, 1980. 4.3.1

[147] D. A. Ausherman, A. Kozma, J. L. Walker, H. M. Jones, and E. C. Poggio, "Developments in radar imaging," *IEEE Trans. Aerospace and Electronic Systems*, vol. 20, pp. 363–400, 1984. 4.3.1

[148] A. Maki and K. Fukui, "Ship identification in sequential ISAR imagery," *Mach. Vision Appl.*, vol. 15, pp. 149–155, 2004. 4.3.1

[149] R. White and M. Williams, "Processing ISAR and spotlight SAR data to very high resolution," in *IEEE International Geoscience and Remote Sensing Symposium*, (Hamburg, Germany), pp. 32–34, 1999. 4.3.2, C.2.1

[150] A. Kovács and T. Szirányi, "Orientation based building outline extraction in aerial images," in *XXII. ISPRS Congress*, vol. I-7 of *ISPRS Annals Photogram. Rem. Sens. and Spat. Inf. Sci.*, pp. 141–146, Melbourne, Australia: ISPRS, 2012. 5.5.1

[151] O. Krammer and B. Sinkovics, "Improved method for determining the shear strength of chip component solder joints," *Microelectronics Reliability*, vol. 50, no. 2, pp. 235 – 241, 2010. 5.5.3

[152] A. Manno-Kovács and A. Ok, "Building detection from monocular VHR images by integrated urban area knowledge," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, pp. 2140–2144, Oct 2015. 5.6

[153] Á. Rakusz, T. Lovas, and Á. Barsi, "Lidar-based vehicle segmentation," in *XX. ISPRS Congress*, vol. XXXV-2 of *ISPRS Archives Photogram. Rem. Sens. and Spat. Inf. Sci.*, pp. 156–159, 2004. 5.7, 5.3

[154] W. Yao, S. Hinz, and U. Stilla, "Automatic vehicle extraction from airborne LiDAR data of urban areas aided by geodesic morphology," *Pattern Recogn. Letters*, vol. 31, no. 10, pp. 1100 – 1108, 2010. 5.7, 5.3

[155] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Computing Surveys*, vol. 38, no. 4, 2006. 6.1

[156] W. Ge and R. T. Collins, "Marked point processes for crowd counting," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Miami, FL, USA), pp. 2913–2920, 2009. 6.1

[157] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 267–282, 2008. 6.1, 6.2.4

[158] W. Ge and R. T. Collins, "Crowd detection with a multiview sampler," in *European Conference on Computer Vision (ECCV)*, vol. 6315 of *Lecture Notes in Computer Science*, pp. 324–337, Heraklion, Crete, Greece: Springer, 2010. 6.1

[159] I. Mikic, P. Cosman, G. Kogut, and M. M. Trivedi, "Moving shadow and object detection in traffic scenes," in *International Conference on Pattern Recognition (ICPR)*, vol. 1, (Barcelona, Spain), pp. 321–324, 2000. 6.1

[160] Y. Wang, K.-F. Loe, and J.-K. Wu, "A dynamic conditional random field model for foreground and shadow segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 279–289, 2006. 6.1, 6.10(c), 6.3.1.3, 6.3.6.1

[161] I. Schiller and R. Koch, "Improved video segmentation by adaptive combination of depth keying and Mixture-of-Gaussians," in *Scandinavian Conference on Image Analysis*, vol. 6688 of *Lecture Notes in Computer Science*, pp. 59–68, Springer, 2011. 6.1

[162] R. Kaestner, N. Engelhard, R. Triebel, and R.Siegwart, "A Bayesian approach to learning 3D representations of dynamic environments," in *International Symposium on Experimental Robotics (ISER)*, (Berlin, Germany), Springer, 2010. 6.1, 6.3.1, 6.10(b), 6.3.1.2, 6.3.6.1

[163] H. Zheng, R. Wang, and S. Xu, "Recognizing street lighting poles from mobile LiDAR data," *IEEE Trans. Geoscience and Remote Sensing*, vol. 55, no. 1, pp. 407–420, 2017. 6.1

[164] Y. Yu, J. Li, H. Guan, and C. Wang, "Automated detection of three-dimensional cars in mobile laser scanning point clouds using DBM-Hough-Forests," *IEEE Trans. Geoscience and Remote Sensing*, vol. 54, no. 7, pp. 4130–4142, 2016. 6.1

[165] B. Wu, B. Yu, W. Yue, S. Shu, W. Tan, C. Hu, Y. Huang, J. Wu, and H. Liu, "A voxel-based method for automated identification and morphological parameters estimation of individual street trees from mobile laser scanning data," *Remote Sensing*, vol. 5, no. 2, p. 584, 2013. 6.1

[166] Y. Wang and T. Tan, "Adaptive foreground and shadow detection in image sequences," in *International Conference on Pattern Recognition (ICPR)*, (Quebec, Canada), pp. 983–986, 2002. 6.3

[167] R. Y. Tsai, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf tv cameras and lenses," *IEEE Journal of Robotics and Automation*, vol. 3, no. 4, pp. 323–344, 1987. 6.2.1

[168] PETS, "Dataset - Performance Evaluation of Tracking and Surveillance," 2009. http://www.cvg.rdg.ac.uk/PETS2009/a.html. 6.2.4

[169] B. Kalyan, K. W. Lee, W. S. Wijesoma, D. Moratuwage, and N. M. Patrikalakis, "A random finite set based detection and tracking using 3D LIDAR in dynamic environments.," in *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, (Istanbul, Turkey), pp. 2288–2292, IEEE, 2010. 6.3.1, 6.3.1.1

[170] N. Muhammad and S. Lacroix, "Calibration of a rotating multi-beam Lidar.," in *International Conference on Intelligent Robots and Systems (IROS)*, (Taipei, Taiwan), pp. 5648–5653, IEEE, 2010. 6.3.1.1

[171] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747–757, 2000. 6.3.1.2, 6.3.1.2, 6.3.6.1

[172] D. Zhang and G. Lu, "A comparative study of fourier descriptors for shape representation and retrieval," in *Asian Conference on Computer Vision (ACCV*, pp. 646–651, Springer, 2002. 6.3.2

[173] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, pp. 316–322, Feb 2006. 6.3.3, 6.3.3, 6.3.3, 6.3.6.2

[174] D. Ciresan, U. Meier, J. Masci, and J. Schmidhuber, "A committee of neural networks for traffic sign classification," in *International Joint Conference on Neural Networks (IJCNN)*, pp. 1918–1921, July 2011. 6.3.3

[175] Y. LeCun, L. Jackel, L. Bottou, A. Brunot, C. Cortes, J. Denker, H. Drucker, I. Guyon, U. Müller, E. Säckinger, P. Simard, and V. Vapnik, "Comparison of learning algorithms for handwritten digit recognition," in *International Conference on Artificial Neural Networks*, (Perth, Australia), pp. 53–60, 1995. 6.3.4

[176] F. Lafarge and C. Mallet, "Creating large-scale city models from 3D-point clouds: A robust approach with hybrid representation," *Int. J. of Computer Vision*, 2012. 6.3.6.1

[177] A. Kale, N. Cuntoor, B. Yegnanarayana, A. Rajagopalan, and R. Chellappa, "Gait analysis for human identification," in *Audio- and Video-Based Biometric Person Authentication*, vol. 2688 of *Lecture Notes in Computer Science*, pp. 706–714, Springer, 2003. 6.3.6.2

[178] M. Hofmann, S. Bachmann, and G. Rigoll, "2.5D gait biometrics using the depth gradient histogram energy image," in *Int'l Conf. on Biometrics: Theory, Applications and Systems (BTAS)*, pp. 399–403, Sept 2012. 6.3.6.2

[179] J. Tang, J. Luo, T. Tjahjadi, and Y. Gao, "2.5D multi-view gait recognition based on point cloud registration," *Sensors*, vol. 14, no. 4, pp. 6124–6143, 2014. 6.3.6.2

[180] A. Azim and O. Aycard, "Detection, classification and tracking of moving objects in a 3D environment.," in *IEEE Intelligent Vehicles Symposium (IV)*, (Alcalá de Henares, Spain), pp. 802–807, 2012. 6.4.2

[181] M. Himmelsbach, A. Müller, T. Luettel, and H.-J. Wuensche, "LIDAR-based 3D Object Perception," in *Proceedings of 1st International Workshop on Cognition for Technical Systems*, (Munich), 2008. 6.4.2

[182] M. De Deuge, A. Quadros, C. Hung, and B. Douillard, "Unsupervised feature learning for outdoor 3D scans," *Proceedings of Australasian Conference on Robotics and Automation*, 2013. 6.4.3, 6.4.7

[183] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. J. Goodfellow, A. Bergeron, N. Bouchard, and Y. Bengio, "Theano: new features and speed improvements," in *Deep Learning and Unsupervised Feature Learning Workshop at NIPS*, (Lake Tahoe, USA), 2012. 6.4.3

[184] Z. Zhang, "Iterative point matching for registration of free-form curves and surfaces," *International Journal of Computer Vision*, vol. 13, no. 2, pp. 119–152, 1994. 6.4.5

[185] M. Magnusson, *The Three-Dimensional Normal-Distributions Transform – an Efficient Representation for Registration, Surface Analysis, and Loop Detection*. PhD thesis, Örebro University, December 2009. 6.4.5, 6.4.5

[186] N. K. Ratha, K. Karu, S. Chen, and A. K. Jain, "A real-time matching system for large fingerprint databases," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 799–813, 1996. 6.4.5

[187] K. Lai and D. Fox, "Object recognition in 3D point clouds using web data and domain adaptation," *Int. J. Rob. Res.*, vol. 29, pp. 1019–1037, July 2010. 6.4.7

[188] R. Rusu and S. Cousins, "3D is here: Point Cloud Library (PCL)," in *IEEE International Conference on Robotics and Automation (ICRA)*, (Shanghai, China), 2011. 6.4.7, 6.4(a), 7.1, E.1

[189] Y. Yu, J. Li, J. Yu, H. Guan, and C. Wang, "Pairwise three-dimensional shape context for partial object matching and retrieval on mobile laser scanning data," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, pp. 1019–1023, May 2014. 6.4.7

[190] J. Huang and S. You, "Point cloud labeling using 3D convolutional neural network," in *International Conference on Pattern Recognition (ICPR)*, (Cancun, Mexico), pp. 2670–2675, 2016. 6.4.7

[191] K. Liu, J. Boehm, and C. Alis, "Change detection of mobile LIDAR data using cloud computing," *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLI-B3, pp. 309–313, Jun 2016. 6.4(b), 6.4.7

[192] R. Laganière, *OpenCV 2 Computer Vision Application Programming Cookbook*. Packt. Pub. Limited, 2011. 7.1

[193] A. Börcs, *Four-dimensional Analysis of Dynamic Urban Environments in Terrestrial and Airborne LiDAR Data*. PhD thesis, Budapest University of Technology and Economics, June 2018. 7.2, 7.2

[194] Y. Sheikh and M. Shah, "Bayesian modeling of dynamic scenes for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 11, pp. 1778–1792, 2005. C.2.1, C.2.1

[195] A. Andrew, "Another efficient algorithm for convex hulls in two dimensions," *Information Processing Letters*, vol. 9, no. 5, pp. 216 – 219, 1979. E

[196] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, (Providence, Rhode Island, USA), pp. 3354–3361, 2012. E.1